Reasoning About Responsibility for Taking Risks

Maksim Gladyshev
Utrecht University
Utrecht, Netherlands
m.gladyshev@uu.nl

Natasha Alechina
Open University
Heerlen, Netherlands
n.a.alechina@uu.nl

Mehdi Dastani
Utrecht University
Utrecht, Netherlands
m.m.dastani@uu.nl

Dragan Doder
Utrecht University
Utrecht, Netherlands
d.doder@uu.nl

ABSTRACT

Tracing responsibility and assigning blame to decision-making actors in multi-agent systems has drawn attention of MAS community in recent years. Existing approaches have proposed several definitions for multi-agent responsibility, but while these definitions differ in details, most of them agree that responsibility for some event may be allocated to agents only if the event is actually realized. In this paper, we argue that in many scenarios this restriction is too strong and that an undesirable outcome may be understood as a (high) risk of some event, and not necessarily as the realization of the event. To cover such cases, we propose a logical framework for reasoning about various notions of multi-agent responsibility for taking risks. The proposed logic contains primitives for probability, strategic power, time and knowledge modalities, which, as we demonstrate, allow to express various notions of group responsibility in probabilistic settings. As the main result we prove that the proposed logic has a complete axiomatization, a decidable satisfiability problem, and an efficient model-checking procedure.

KEYWORDS

Responsibility; Logics for MAS; Probabilistic Logic

ACM Reference Format:
Maksim Gladyshev, Natasha Alechina, Mehdi Dastani, and Dragan Doder. 2026. Reasoning About Responsibility for Taking Risks. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/OAJK8159

1 INTRODUCTION

With the rapid development of complex multi-agent systems, there is obvious urgency for ensuring their reliability, safety and trustworthiness [5, 8, 23, 34]. If an AI system creates an unsafe outcome, it is crucial to determine which part of the system is responsible for this event. This urgently calls for the development of formal tools and reasoning methods to identify the responsibility of agents whose actions lead to the unsafe outcomes [8, 33, 38].

The recently proposed approaches in the AI literature study responsibility from various perspectives, such as Strategic reasoning [3, 7, 24, 25, 30, 36, 37], Stit-style logics, [21, 26], planning [1, 28], game-theory [2, 32] and causal reasoning [4, 17]. The existing literature studies the notion of responsibility from different angles. The first important distinction is active vs passive responsibility. Active responsibility is attributed to the agents when their actions make an outcome unavoidable, while passive responsibility is a counterfactual notion applicable when an undesirable outcome could have been prevented by the agents. Another important distinction is between *ex ante* and *ex post* (i.e. forward- and backward-looking) approaches. *Ex ante* responsibility is considered when the agents *may* become responsible by performing certain actions, while *ex post* responsibility can be attributed to the agents only after their actions are executed and the result of their choice is known. In this paper we focus on *passive* and *ex post* responsibility.

Despite these differences, most of existing approaches assume that in order to claim that (a group of) agent(s) is responsible for an undesirable outcome, this outcome must be realized. In this paper we take a different approach and study the notion of responsibility that covers situations in which a harmful outcome is not necessarily realized. In particular, we consider situations where agents can be blamed because they have created a substantial risk of occurrence of a harmful outcome, while they had more prudent options. This approach is reflected in various endangerment offences like leaving firearms accessible to a child or various forms of reckless driving.

The general notion of (counterfactual) group *responsibility* states that a group of agents G is *responsible* for some outcome φ if (R1) φ actually takes place; (R2) group G had a strategy (intervention) to prevent φ in the past; (R3) group G is minimal, i.e., no proper subset of G satisfies (R2). Alternative notions of responsibility differ in details, but pretty much follow this structure. Note that, in case of agents' uncertainty about their current state (i.e. partial observability), the existence of a strategy/intervention to prevent the harmful outcome does not guarantee agents' awareness about such strategy or intervention. For example, knowing that there is an action (correct password) to open a safe does not mean knowing which action can open it (which password is correct). To cover such cases, (R2) condition should additionally require that the agents in G not only have a strategy to prevent φ, but also *know* which strategy prevents φ. Such strategies are also known as uniform as they select one and the same action in indistinguishable states. This general structure that defines the notion of strategic responsibility appears in various works on the topic, e.g. [3, 7, 24, 26, 36, 37].

We extend this framework for reasoning about multi-agent responsibility to probabilistic settings, where the *risk of* an event

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/OAJK8159

is taken into account, not only the *realization of* the event. On the conceptual side, we deem our main contribution in the new notions of responsibility for probabilistic contexts, which take into account how different choices of agents affect the probability of occurrence of a negative outcome, and the formal framework suitable for representing these notions. On the technical side, we develop a rich logical framework with several modalities, in which we can represent our notions of responsibility for probabilistic settings and reason about them. Our logic contains the operator $[\delta_G]$, which allows us to refer to a particular action profile δ_G of a group G. This operator was previously used for reasoning about responsibility in [7, 30]. A logic of the operator $[\delta]$ is called Coalition Epistemic Dynamic Logic (CEDL) due to its resemblance to Propositional Dynamic Logic (PDL) [13]. Together with a standard modal operator for knowledge, it allows us to express the existence of a uniform strategy for agents in G. We also use a temporal operator Y to refer to the past state. This combination already allows us to express responsibility in deterministic settings. However, because our main goal is reasoning about responsibility in probabilistic context, we also introduce a modal operator $Pr() \geq \alpha$ for probabilities. We present a sound and complete axiomatization for the logic, along with an efficient model checking algorithm.

This paper is organized as follows. In Section 2 we briefly discuss the motivation behind new definitions on a few informal examples. In Section 3 we formally define our class of models, introduce our logic and discuss our modelling choices. In Section 4 we propose new notions of responsibility suitable for probabilistic contexts. Finally, in Section 5 we present a completeness proof for our logic and demonstrate that it has a decidable satisfiability problem and an efficient model-checking procedure. We conclude in Section 6.

2 MOTIVATIONAL EXAMPLES

To highlight the drawbacks of the deterministic interpretation of responsibility, consider two toy examples.

Example 1. *Assume that h is a health care advisor agent prescribing a treatment to a patient. Assume that in the initial state, without a treatment the patient dies with the probability of* 60%. *Our agent has three choices of the treatment A, B and C, that can still result with death of the patient with probability of* 55%, 30% *and* 10%, *respectively.*

It is easy to see that the notion of non-probabilistic responsibility mentioned in the Introduction is not suitable for this example. The agent h has no action that can surely prevent the negative outcome, only one that can alter its probability. Then h does not satisfy R2 condition and thus cannot be claimed responsible. However, if h chooses either A or B, we believe it is reasonable to consider h blameworthy for not making the probability of the patient's death *as low as possible*. More generally, in such probabilistic settings, we consider a group of agents G to be responsible for *not choosing the safest option* in the initial state. This is the first notion of responsibility that we aim to formalize in this paper.

Though this interpretation may be crucial in many risk-sensitive domains, in some applications choosing the safest option is not always required or desirable. Consider the following example:

Example 2. *Consider a self-driving car c moving along the highway. Assume that under normal conditions a risk of an accident is .001%. The car can maintain its current speed (M) (so the risk remains .001%), increase the speed (I) (increasing the risk to .005%) or stop at a parking lot (S) reducing the risk to 0.*

According to the above-mentioned notion of responsibility in probabilistic setting, only action (S) would be considered admissible in this situation, because (S) is the safest option. Moreover, if the agent chooses to maintain its speed and gets into an accident, even a deterministic definition would claim the agent's responsibility, because a negative outcome is realized (R1) and c could prevent it by choosing (S) in the initial state (R2). However, in scenarios such as Example 2 we may not want to claim that agents are responsible as long as they do not *increase* the current risk of a negative outcome. This interpretation gives rise to another notion of being responsible for increasing risk.

We believe that these examples demonstrate that deterministic view on multi-agent responsibility is too limited for many scenarios. This advocates the need for formal tools to represent similar examples and reason about responsibility allocation in non-deterministic setting.

3 FORMAL FRAMEWORK

We begin by introducing the class of our models and illustrating them with examples. Then, we define a logic for reasoning about these structures. Let us fix a finite set of agents $\mathbb{AG} = \{a_1, \ldots, a_n\}$, a finite set of atomic actions Act and a countable set of propositional variables $Prop$. Let $F \subseteq [0,1] \cap \mathbb{Q}$ denote a *finite* subset of rational numbers from $[0,1]$, such that $\{0,1\} \subseteq F$. We assume that F contains finitely many possible values for our probability function. Moreover, we restrict our models to be tree-like such that each state has at most one predecessor.

Our models are essentially Epistemic Concurrent Game Structures (ECGS) endowed with probability functions.

Definition 1 (Models). *A* model *is a tuple*

$$M = (S, S_f, \{\sim_i\}_{i \in \mathbb{AG}}, act, o, P, L)$$

where

- S *is a nonempty finite set of states and* $S_f \subseteq S$ *is a set of* final *states.*
- $\sim_i \subseteq S \times S$ *is an equivalence (epistemic) relation for all* $i \in \mathbb{AG}$.
- *Availability function* $act : \mathbb{AG} \times S \longrightarrow 2^{Act} \setminus \{\emptyset\}$ *defines nonempty sets of actions available to agents at each state.*
- o *is a* non-deterministic *transition function that assigns the outcome states* $X = o(s, (\delta_1, \ldots, \delta_n))$ *(where* $X \subseteq S$*) to a state* $s \in \overline{S_f}$ *and a tuple of actions* $(\delta_1, \ldots, \delta_n)$ *with* $\delta_i \in act(i, s)$, *that can be executed by* $\mathbb{AG}$ *in* s. *To impose the restriction that each* $s \in S$ *has at most one predecessor, we require* $\forall s \in S, |o^-(s)| \leq 1$, *where* $o^-(s)$ *denotes the set of states from which a transition to* s *is possible, i.e.,* $s' \in o^-(s)$ *if* $\exists \delta : s \in o(s', \delta)$.
- $P : S \mapsto (2^S \mapsto F)$ *assigns a probability measure* $P(s)$ *to each state* s. *We require* P *to satisfy the following properties:*

  P1 $P(s)(S) = 1$,

  P2 $P(s)(\emptyset) = 0$,

  P3 $P(s)$ *is (finitely) additive:*

$$P(s)(\bigcup_{0 \leq i \leq m} X_i) = \sum_{0 \leq i \leq m} P(s)(X_i),$$

whenever $X_i \cap X_j = \emptyset$ *for any* $i \neq j$,
P4 $P(s)(\{s\}) > 0$,
P5 $P(s)(\{s'\}) > 0$ *implies* $P(s) = P(s')$.

- $L : Prop \cup \{final\} \to 2^S$ *is a labelling function, such that* $L(final) = S_f$.
- *For any* $\delta_1, \ldots, \delta_n$ *and any* $s \in S$, *if* $X = \{s' \mid s' \in o(s, (\delta_1, \ldots, \delta_n))\}$, *then for any* $s', s'' \in X$
  C1 $P(s')(\{s''\}) > 0$,
  C2 $P(s')(X) = 1$.

We call a group of agents $H \subseteq \mathbb{AG}$ a coalition. A group of all agents $\mathbb{AG}$ is called the grand coalition. For δ_G, which is an action profile of a non-grand coalition $G \subset \mathbb{AG}$, $o(s, \delta_G)$ is defined as the set containing all outcomes of δ_G completed by actions of agents outside the coalition. Δ denotes a set of all possible complete action profiles. We use δ instead of $\delta_{\mathbb{AG}}$ to denote complete action profiles. Sometimes we write $\delta|_G$ to denote δ's restriction to agents in G.

Let us have a closer look at our modelling choices. Our models are based on Concurrent Game Structures endowed with an epistemic uncertainty relation $\sim_i$ for each $i \in \mathbb{AG}$, also known as Epistemic CGSs. However, they have several notable differences. First of all, because we are interested in reasoning in probabilistic settings, the transition function o is non-deterministic. This is a fairly common assumption in the field. Secondly, our models have designated final states $S_f \subseteq S$, for which no outgoing transition is defined. So, those states are deadlocks and violate so-called 'activity' property. And because each state has at most one predecessor, our models represent linear past and branching transitions to reason about counterfactual scenarios.

Each (available) complete action profile δ non-deterministically leads to an element of $X_\delta \subseteq S$. To represent the probabilities of these transitions we use a probability function P. P assigns each state $s \in S$ with a probability measure $P(s)$ on 2^S. Intuitively, if s is the current state after transitioning by δ, $P(s)(\{s_1, s_2\})$ is the probability of ending up in one of the states from $\{s_1, s_2\}$ rather than in s. For $p \in Prop$, sometimes we write $P(s)(\|p\|)$ to refer to $P(s)(\{s' \mid p \in s\})$.

Note that the way we model probabilities of transitions is different from modelling in stochastic games. In a stochastic game, transition probabilities are associated with *actions* and they are 'forward-looking': being in a state s, $P_\delta(s)(s_1)$ denotes probability that the next transition after executing δ in s will lead to s_1; while our probabilities are associated with *states* and are 'backward-looking': being in a state s, $P(s)(s_1)$ denotes probability that we could have ended up in s_1 (instead of s) after the previous transition. Our models have linear past: being in a state s we always know which state $Past(s)$ was the previous one (and thus we know which action was executed there). This is not true for an arbitrary stochastic game: different states may be potential predecessors of the current state s, which can create difficulties for responsibility allocation. However, our models may be viewed as unfoldings of stochastic games up to some finite depth.

The conditions $(P1) - (P3)$ guarantee that $P(s)$ is a probability measure. In order to explain the intuition behind our probabilities and conditions $(P4) - (P5)$ and $(C1) - (C2)$, we illustrate our definition employing Example 1 from the previous section.

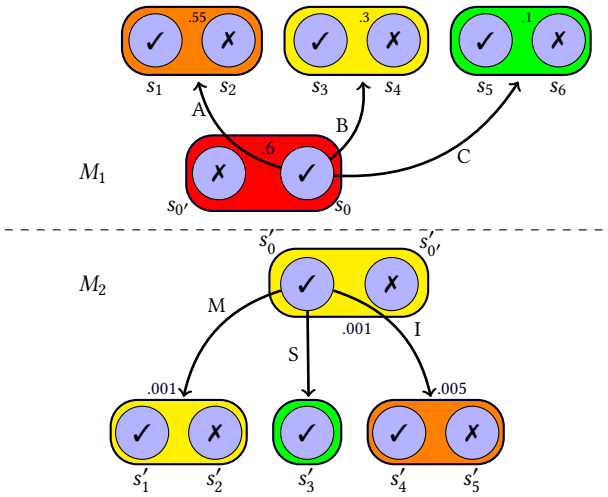

Figure 1: Models M_1 and M_2 for Examples 1 and 2. Here the circles represent the states, ✗ denotes states at which a negative outcome (ie., "patient dies" for M_1, and "car accident" for M_2) has happened, and ✓denotes 'safe' states. Rectangles denote support sets of different probability distributions $P(s)$. Due to P5, states s, s' from the same support set ($P(s)(s') > 0$) have the same probability distribution ($P(s) = P(s')$). Different colours also denote different probabilities of ✗. Arrows marked with action profiles represent the transition function o. The states without an outgoing transition are *final*.

Continuance of Example 1. *Let us introduce the model M_1, depicted in the upper part of Figure 1, which models the scenario with three treatments with different death rates. In (M_1, s_0) treatment A leads to either the state s_1, in which the patient survives, or to s_2, in which the patient dies (denoted by the propositional letter $p_✗$). Thus, $o(s_0, A) = \{s_1, s_2\}$. In addition to information about death ($p_✗ \notin s_1$, $p_✗ \in s_2$), both states also contain information about alternate possibilities after executing the action A (according to o), with probabilities in ending in each of them. In particular, the fact that the chance of death was .55 is modelled by $P(s_1)(s_2) = P(s_2)(s_2) = .55$. Consequently, the complementary chance of survival is $P(s_1)(s_1) = P(s_2)(s_1) = .45$ (we omit those complementary values from Figure 1 for clarity).*

Note also that $P(s_1)(\{s_1, s_2\}) = P(s_2)(\{s_1, s_2\}) = 1$, so for any $s' \in S - \{s_1, s_2\}, P(s_1)(s') = 0$. Recall that due to (C1) and (P5), $P(s_1)(s') = 0$ implies $P(s_2)(s') = 0$. The alternative actions (B) and (C) lead to states with different probabilities: (B) leads to $\{s_3, s_4\}$ with $P(s_3)(\|p_✗\|) = P(s_3)(s_4) = .3$ and $P(s_4)(s_4) = .3$, and (C) leads to $\{s_5, s_6\}$ with $P(s_5)(\|p_✗\|) = P(s_5)(s_6) = .1$ and $P(s_6)(s_6) = .1$. In the initial state s_0, $P(s_0)(\|p_✗\|) = .6$.

As illustrated above, $P(s)(s') > 0$ means that s and s' are two possible outcomes of the same complete action profile δ, and the probability of s' being the outcome of δ is $P(s)(s')$.

Thus, each state must assign itself non-zero probability, which is guaranteed by P4. P5 ensures that given s all states $s' \in \{s'' \in S \mid P(s)(s'') > 0\}$ in its support set have the same probability distributions $P(s) = P(s')$. These two properties correspond to

S5 conditions for epistemic logic with uncertainty represented by probability measure [9, 19].

Finally, C1 and C2 ensure that the outcomes of a non-deterministic transition function o all agree on the same probability measure P, in the way that the probability of each state corresponds to the probability of transition resulting in that outcome, and the set of all possible outcomes of a transition has probability 1.

Now we also provide a model for the second motivating example from the previous section.

Continuance of Example 2. *The model M_2, depicted in the Figure 1, models the scenario from the Example 2. Here $p_{\text{✗}}$ denotes the accident. Similarly as above, in (M_2, s'_0) it holds that $P(s'_0)(\|p_{\text{✗}}\|) = .001$, action (M) leads to $\{s'_1, s'_2\}$ with $P(s'_1)(\|p_{\text{✗}}\|) = .001$, action (S) (deterministically) leads to $\{s'_3\}$ with $P(s'_3)(\|p_{\text{✗}}\|) = 0$, and action (I) leads to $\{s'_4, s'_5\}$ with $P(s'_4)(\|p_{\text{✗}}\|) = .005$.*

We would like to emphasize that in our settings values of $P(s)$ belong to a finite set F, instead of a real interval $[0, 1]$ as typical in probabilistic modal logic [10, 19]. This restriction to a finate set of values was previously explored in [12] and [35]. On the one hand, this restriction provides more technical convenience especially in cases of quantification over F instead of over $\mathbb{R} \cap [0, 1]$ as discussed in Section 6. On the other hand, it is usually sufficient for modelling and verification purposes to have a finite representation of the model anyway. That is why we believe the proposed restriction is suitable for our purposes.

Note also that we do not impose additional restrictions on epistemic relation $\sim_i$. However, in some scenarios additional interactions between epistemic relations and transition function can be introduced. For example, one may require that agents are always aware about actions available to them: $s_1 \sim_i s_2$ implies $act(i, s_1) = act(i, s_2)$, or that agents have 'perfect recall' and never loose information: $s_1 \in o^-(s_2)$ implies $\sim_i (s_2) \subseteq \sim_i (s_1)$ for any $i \in \mathbb{AG}$ [7, 30].

3.1 Language and Semantics

Definition 2 (Language). *The language $\mathcal{L}$ of our logic is defined recursively by the following Backus–Naur form:*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid Pr(\varphi) \geq \alpha \mid [\delta_G]\varphi \mid \mathsf{Y}\varphi \mid K_i\varphi,$$

where $p \in Prop$, α is any rational in $[0, 1]$ and $\delta_G = \delta|_G$ for some $\delta \in \Delta$. Boolean connectives are defined in a usual way. Operator $Pr(\varphi) \geq \alpha$ means "probability of φ is at least α". Derived operators $Pr(\varphi) \bowtie \alpha$ for $\bowtie \in \{>, =, \leq, <\}$ are introduced in the following way: $Pr(\varphi) < \alpha =_{def} \neg Pr(\varphi) \geq \alpha$, $Pr(\varphi) \leq \alpha =_{def} Pr(\neg\varphi) \geq 1 - \alpha$, $Pr(\varphi) > \alpha =_{def} \neg Pr(\varphi) \leq \alpha$, $Pr(\varphi) = \alpha =_{def} Pr(\varphi) \geq \alpha \wedge Pr(\varphi) \leq \alpha$. Operator $[\delta_G]\varphi$ means "after every execution of actions δ_G, φ is true". The dual $\langle\delta_G\rangle = \neg[\delta_G]\neg$ means "there exists an execution δ', s.t. $\delta'|_G = \delta_G$, after which φ is true". Operator $\mathsf{Y}\varphi$ means "φ was true on the previous step (Yesterday)". $K_i\varphi$ means "agent i knows φ". Abbreviation $E_G\varphi = \bigwedge_{i\in G} K_i\varphi$ denotes "everybody in G knows φ".

Firstly, note that probability values α in the formulae are not restricted to the finite range F and can be any rational in $[0, 1]$. Secondly, note that due to the finite nature of both probability range F and the set of complete action profiles Δ, we can quantify over those sets using finite conjunctions and disjunctions. Thus, we will simplify the notation and use $\forall_{r\in F}$ instead of $\bigwedge_{r\in F}$, $\exists_{r\in F}$ instead of $\bigvee_{r\in F}$, and $\forall_\delta$ instead of $\bigwedge_{\delta\in\Delta}$, $\exists_\delta$ instead of $\bigvee_{\delta\in\Delta}$.

Definition 3 (Semantics). *Given a model M, a state $s \in S$ and a formula $\varphi \in \mathcal{L}$ we define the satisfaction relation $\Vdash$ in the following way (the cases for p and boolean connectives are defined standardly):*

- $(M, s) \Vdash Pr(\varphi) \geq \alpha$ *iff* $P(s)(\|\varphi\|^M) \geq \alpha$, *where* $(\|\varphi\|^M) = \{s \in S \mid (M, s) \Vdash \varphi\}$;
- $(M, s) \Vdash [\delta_G]\varphi$ *iff* $\forall s' \in o(s, \delta_G)$ *it holds that* $(M, s') \Vdash \varphi$;
- $(M, s) \Vdash \mathsf{Y}\varphi$ *iff* $\exists s' \in o^-(s) : (M, s) \Vdash \varphi$;
- $(M, s) \Vdash K_i\varphi$ *iff* $s \sim_i s'$ *implies* $(M, s') \Vdash \varphi$.

We also write $M \Vdash \varphi$ if $(M, s) \Vdash \varphi$ for all $s \in S$, and $\Vdash \varphi$ if $M \Vdash \varphi$ for any M. In the latter case we call φ *valid*.

When we write $\delta_{i:a}$ we mean that agent i takes action a. The following statements about Figure 1 illustrate the semantics.
$(M_1, s_1) \Vdash p_{\text{✓}} \wedge Pr(p_{\text{✗}}) = .55 \wedge \mathsf{Y}K_h[\delta_{h:C}]Pr(p_{\text{✗}}) = .1$
states that "in (M_1, s_1) : $p_{\text{✓}}$ is true, probability of $p_{\text{✗}}$ is .55, and on the previous step h knew that using action C will enforce the probability of $p_{\text{✗}}$ to be .1", and
$(M_2, s_2) \Vdash p_{\text{✗}} \wedge Pr(p_{\text{✗}}) = .001 \wedge \mathsf{Y}\exists\delta_c[\delta_c](p_{\text{✓}} \wedge Pr(p_{\text{✗}}) = 0)$ states that "in (M_2, s_2) the negative outcome is true (an accident happens), its probability is .001, and on the previous step agent c had a strategy to make the probability to be 0 and avoid the negative outcome".

It is easy to check that these formulas hold in Figure 1 according to Definition 3.

4 RESPONSIBILITY FOR TAKING RISKS

As we discussed in the Introduction, in the simplest deterministic form, the notion of counterfactual responsibility may be formulated as follows. A group G is responsible for φ at (M, s) if (R1) φ holds in (M, s), (R2) there exists a G-strategy δ_G, such that group G knows that δ_G enforces $\neg\varphi$, and (R3) no proper subset of G satisfies (R2). This notion translates to our language as

$$Resp_G(\varphi) \equiv \varphi \wedge \mathsf{Y}\exists_{\delta_G} E_G[\delta_G]\neg\varphi \wedge \neg\exists_{H\subset G}\mathsf{Y}\exists_{\delta_H} E_H[\delta_H]\neg\varphi$$

Similar definitions[1] appear in various papers on the topic, e.g. [3, 24, 30]. However, in our examples from Section 2 this definition does not behave in a desirable way. In Example 1, $M_1 \nVdash Resp_h(p_{\text{✗}})$, so no state of M_1 satisfies the definition above. In Example 2 the situation is more interesting, because both (M_2, s'_2) and (M_2, s'_5) satisfy the definition. Thus, in both states (R1) $p_{\text{✗}}$ is satisfied, (R2) c has a strategy (S) that enforces $\neg p_{\text{✗}}$ (and c knows about this), and (R3) $\{h\}$ is a singleton, so the minimality condition is satisfied. We believe this behavior of $Resp_G(\varphi)$ can be considered undesirable by the following reason. First of all, the definition takes into account only the truth value of $p_{\text{✗}}$, but not its probability. Thus, both states s'_4 and s'_5 agree that the probability of $p_{\text{✗}}$ is .005, however $(M_2, s'_4) \nVdash Resp_c p_{\text{✗}}$ (because (R1) condition is violated) while $(M_2, s'_5) \Vdash Resp_c p_{\text{✗}}$. Secondly, as we already discussed, depending on a situation the safest action is not necessarily an optimal choice.

Recently, a notion of responsibility for risky behaviour was proposed in [14, 15]. According to this notion, a group of agents is held responsible if the probability of a harmful event is above a

[1] Note that here we used 'everybody knows' modality E_G to define responsibility, instead of distributed or common knowledge [11]. We believe that distributed knowledge is too weak to claim agents blameworthy, while common knowledge is excessively strong to be applicable. However, both notions can be implemented in our framework.

fixed numerical threshold α, while they could take actions that would push the probability of the event below that threshold. This definition can be encoded in our language as

$$Resp_G(\varphi, \alpha) \equiv Pr(\varphi) > \alpha \wedge \mathsf{Y}\exists_{\delta_G} E_G[\delta_G]Pr(\varphi) \leq \alpha \\ \wedge \forall_{H\subset G}\mathsf{Y}\neg\exists_{\delta_G} E_G[\delta_G]Pr(\varphi) \leq \alpha$$

Note that in [14] only perfect information scenarios were considered, while our framework allows to take into account agents' uncertainty about the current state.

To illustrate the intuition, let us fix a risk threshold $\alpha = .001$. Then, in Figure 1 for $s' \in \{s'_4, s'_5\}$ it holds that $(M_2, s') \Vdash Resp_c(p_{\text{✗}}, \alpha)$. However, if we increase the threshold to $\alpha' = .005$, then $M_2 \nVdash Resp_c(p_{\text{✗}}, \alpha')$, because risk of an accident in s'_4 and s'_5 does not exceed the threshold.

While being useful in many applications, we believe this approach is limited as legal doctrines and endangerment offences usually do not specify a fixed probability value which would separate responsible from irresponsible behaviour. And as we argued in Section 2, often additional definitions are needed. Thus, the idea of being responsible for *not keeping the probability of φ as low as possible* may be expressed as

$$Resp^{min}_G(\varphi) \equiv \exists_{r\in F}(Pr(\varphi) \geq r \wedge \mathsf{Y}(\exists_{\delta_G} E_G[\delta_G]Pr(\varphi) < r \\ \wedge \forall_{H\subset G}\neg\exists_{\delta_H} E_H[\delta_H]Pr(\varphi) < r))$$

In Example 1, $Resp^{min}_h(p_{\text{✗}})$ holds at states $\{s_1, s_2, s_3, s_4\}$. So, according to this definition, agent h is considered blameworthy if the chosen action is (A) or (B), because in this case the safest option (C) is ignored. However, in Example 2 not only states $\{s'_4, s'_5\}$ satisfy $Resp^{min}_c(p_{\text{✗}})$, but also states s'_1, s'_2, because decision to maintain the current speed (M) does not minimize the risk of an accident. In order to encode the intuition behind Example 2, we introduce a final notion of responsibility for *increasing the risk* that can be formulated as follows:

$$Resp^{\uparrow}_G(\varphi) \equiv \exists_{r\in F}\Big(Pr(\varphi) > r \wedge \mathsf{Y}\big(Pr(\varphi) = r \wedge \\ \exists_{\delta_G} E_G[\delta_G]Pr(\varphi) \leq r) \wedge \forall_{H\subset G}\neg\exists_{\delta_H} E_H[\delta_H]Pr(\varphi) \leq r\big)\Big)$$

Now, $Resp^{\uparrow}_c(p_{\text{✗}})$ is satisfied only in states $\{s'_4, s'_5\}$ of M_2 in Figure 1, because only action (I) (increase the speed) *increases* the risk of $p_{\text{✗}}$ with respect to the initial state. However, because in M_1 no action of agent h increases the risk wrt s_0 it holds that $M_1 \nVdash Resp^{\uparrow}_h(p_{\text{✗}})$.

Although for simplicity of the presentation all our examples illustrate one step interaction of agents, and we only demonstrated how to encode single step definitions of responsibility, the modular nature of the proposed framework allows reasoning about multiple step interactions as well. For example, instead of 1-step clause $\mathsf{Y}\exists_{\delta_G} E_G[\delta_G]\neg\varphi$, meaning that on the previous step G had a strategy δ_G such that G knew that δ_G prevents φ, one might define 2-steps version of this clause by nesting past-time and strategic ability modalities as $\mathsf{YY}\exists_{\delta_G} E_G[\delta_G]\exists_{\delta'_G} E_G[\delta'_G]\neg\varphi$. So, by nesting different modalities in our framework one may obtain a definition of responsibility for any fixed temporal depth. At the same time, we admit that our logic is still unable to capture LTL-style modality 'eventually in the past' that does not require fixed temporal depth. Such a modality, together with Probabilistic ATL strategic operators would result in a more expressive framework. However, axiomatizing PATL-style framework is an extremely challenging task. So, we leave it for future work.

To sum up, we believe that alternative modelling settings may require alternative definitions of what it means to be responsible or blameworthy. In some scenarios, we might want agents to stick to the safest strategies and minimize the risk whenever it is possible. However, in different settings, we might be interested in preventing agents from increasing the current risk of some outcome, but not necessarily from minimizing it, or we might have a specific risk threshold that agents are not allowed to cross. Sometimes we are interested in one-shot interactions of agents only, but sometimes we need to deal with temporally extended goals, and then multiple-step definition of responsibility is required. So, we would like to emphasize, that we deem our main contribution to be the proposed logic rather than any particular definitions of responsibility related to risk. And we argue that our logic is expressive enough to encode various such definitions. In the paper we discuss some of them, but, depending on the modelling settings, others may be considered as well. In the next section we demonstrate that the proposed logic admits complete axiomatization, decidable satisfiability problem and efficient model-cheking procedure.

5 AXIOMATIZATION

Let us fix enumeration of $F = \{r_0, \dots, r_n\}$, s.t. $0 = r_0 < \dots < r_i < r_{i+1} < \dots < r_n = 1$. The proof system for our logic L is presented in Table 1. The first group of axioms defines properties of $[\delta_G]$ operator. (AAK) is a standard K axiom for modal logic, (AA1) – (AA4) define the properties of the transition function, and (AA5) – (AA6) encode properties C1 and C2 from Definition 1. The second group of axioms encodes all properties of probabilistic operator $Pr()$. Axioms (PA1) – (PA5) are standard properties of probabilities, (PA6) restricts probabilistic values to finite range F, and (PA_T) – (P_{45}) encode properties P4 and P5 from Definition 1. The next group of axioms deals with past-time operator Y. (YAK) is a standard K axiom, (YA1) enforces uniqueness of the past, and (YA2) – (YA3) encode the interplay of $[\delta]$ and Y operators. Finally, (KAK)-(KA5) are standard S5 axioms for multi-agent epistemic logic. It is a routine to show that all axioms in Table 1 are *sound* (i.e. valid on our class of models), so for space reasons we omit the proof.

A formula φ is a *theorem*, denoted by $\vdash \varphi$, if there is a sequence of formulas $\varphi_0, \varphi, \dots, \varphi_n$, such that $\varphi_n = \varphi$ and every φ_i is an axiom, or it is derived from the preceding formulas by an inference rule. We also say that φ is deducible from a set of formulas T, denoted by $T \vdash \varphi$ if there is a sequence of formulas $\varphi_0, \varphi, \dots, \varphi_n$, such that $\varphi_n = \varphi$ and every φ_i is an theorem, or it is derived from the preceding formulas by **(MP)**.

In this section we prove that the presented axiomatization is complete with respect to the proposed class of models. The proof is organized as follows. At first, we define a sub-language for any consistent formula φ. We call this construction a closure $cl(\varphi)$, due to its resemblance to Fisher-Ladner style closure [13]. Then, we construct a canonical pseudo model $\mathfrak{M}$ over the set of maximally consistent subsets of $cl(\varphi)$. In this pseudo-model, transitions are represented as a Kripke-style binary relations and probability function P is not yet defined. Then, we extend this pseudo model to a

(Taut) All propositional tautologies
Action Axioms:
(AAK) $([\delta_G]\varphi \wedge [\delta_G]\psi) \to [\delta_G](\varphi \wedge \psi)$
(AA1) $\neg final \to \exists_{\delta \in \Delta}\langle\delta\rangle\top$
(AA2) $\langle\delta\rangle\top \to \bigwedge_{i \in \mathbb{AG}}\langle\delta|_i\rangle\top$
(AA3) $(\langle\delta_G\rangle\top \wedge \langle\delta'_H\rangle\top) \to \langle\delta_G \cup \delta'_H\rangle\top$ for $G \cap H = \emptyset$
(AA4) $([\delta_G]\varphi \wedge [\delta'_H]\psi) \to [\delta_G \cup \delta'_H](\varphi \wedge \psi)$ for $G \cap H = \emptyset$
(AA5) $\langle\delta\rangle Pr(\varphi) \bowtie \alpha \to [\delta] Pr(\varphi) \bowtie \alpha$ for $\bowtie \in \{\geq, >, =, \leq, <\}, \alpha \in [0,1]$
(AA6) $[\delta]\varphi \to [\delta] Pr(\varphi) \geq 1$
Probability Axioms:
(PA1) $Pr(\varphi \to \psi) \geq 1 \to ((Pr(\varphi) > \alpha \to Pr(\psi) > \alpha) \wedge (Pr(\varphi) > \alpha \to Pr(\psi) \geq \alpha) \wedge (Pr(\varphi) \geq \alpha \to Pr(\psi) \geq \alpha))$
(PA2) $Pr(\varphi \to \psi) \geq 1 \to (Pr(\varphi) \geq \alpha \to Pr(\psi) \geq \beta)$, $\alpha, \beta \in [0,1], \beta < \alpha$
(PA3) $Pr(\varphi) \geq 0$
(PA4) $Pr(\varphi \vee \psi) > (\alpha + \beta) \to (Pr(\varphi) > \alpha \vee Pr(\psi) > \beta)$, $\alpha, \beta, (\alpha + \beta) \in [0,1]$
(PA5) $Pr(\neg(\varphi \wedge \psi)) \geq 1 \to ((Pr(\varphi) > \alpha \wedge Pr(\psi) \geq \beta) \to \to Pr(\varphi \vee \psi) > (\alpha + \beta))$, $\alpha, \beta, (\alpha + \beta) \in [0,1]$
(PA6) $Pr(\varphi) > r_i \to Pr(\varphi) \geq r_{i+1}$, for $r_i, r_{i+1} \in F$
(PA_T) $Pr(\varphi) \geq 1 \to \varphi$
(PA_{45}) $Pr(\varphi) \bowtie \alpha \to Pr(Pr(\varphi) \bowtie \alpha) \geq 1$
'Yesterday' Axioms:
(YAK) $(\neg Y\neg\varphi \wedge \neg Y\neg\psi) \to \neg Y\neg(\varphi \wedge \psi)$
(YA1) $Y\varphi \to \neg Y\neg\varphi$
(YA2) $\varphi \to [\delta] Y\varphi$
(YA3) $(\varphi \wedge Y\top) \to \exists_\delta Y\langle\delta\rangle\varphi$
Knowledge Axioms:
(KAK) $(K_i\varphi \wedge K_i\psi) \to K_i(\varphi \wedge \psi)$
(KAT) $K_i\varphi \to \varphi$
(KA4) $K_i\varphi \to K_iK_i\varphi$
(KA5) $\neg K_i\varphi \to K_i\neg K_i\varphi$
Inference Rules:
(MP) From φ and $\varphi \to \psi$, infer ψ
(Nec_1) from φ infer $[\delta|_G]\varphi$
(Nec_2) from φ infer $Pr(\varphi) \geq 1$
(Nec_3) from φ, infer $K_i\varphi$
(Nec_4) from φ infer $\neg Y\neg\varphi$

Table 1: Axiomatization AX of our logic L.

proper canonical model M^c, demonstrate that it satisfies all necessary constrains and prove the Truth lemma. In general, our proof uses fairly standard techniques, and the main source of difficulties is a fusion of different modalities and their interplay.

Given an L-consistent formula φ, let $A(\varphi)$ be an ordered set of all rational numbers $\frac{p}{q} \in [0,1]$ where q is the smallest common denominator of all indexes α in φ together with all elements of F.

Definition 4 (Closure). *For any* L*-consistent formula* φ*,* $cl(\varphi)$ *is the smallest set containing all possible formulas* ψ *with* $|\psi| \leq |\varphi| + 1$*, such that*

- *only propositional variables from* φ *and* $\{final\}$ *occur in* $cl(\varphi)$*;*
- *indexes* α *in probabilistic operator belong to the finite set* $A(\varphi)$*;*

Let Ω be the set of all maximally consistent subsets of $cl(\varphi)$. The standard properties of such construction[19, 20] are: (1) Ω is finite; (2) any subset of Ω is an equivalence class of some formula from $cl(\varphi)$ i.e., $\forall X \subseteq \Omega \exists \psi \in cl(\varphi) : X = \|\psi\|$, where $\|\psi\| = \{s \in \Omega \mid \psi \in s\}$; (3) $\|\varphi_1\| \subseteq \|\varphi_2\|$ iff $\vdash \varphi_1 \to \varphi_2$.

The second property is especially useful, because it will allow us to refer to arbitrary states or sets of states in the canonical model using formulas from $cl(\varphi)$. So, given $s \in S$ and $X \subseteq S$ we denote their characteristic formulas as φ_s (s.t. $\|\varphi_s\| = s$) and φ_X (s.t. $\|\varphi_X\| = X$).

At first, consider a pseudo-model defined as follows

Definition 5 (Canonical Pseudo Model). *A canonical pseudo model is the tuple* $\mathfrak{M} = (S^c, S^c_f, \sim^c, \tau, L^c)$*, where*

- $S^c = \Omega$,
- $S^c_f = \{s \in S^c \mid final \in s\}$,
- $\forall s, s' \in S^c : s \sim^c_i s'$ *iff* $\forall K_i\psi \in cl(\varphi) : K_i\psi \in s$ *iff* $K_i\psi \in s'$,
- τ *represents a set of binary relations indexed by all elements of* $act^{\mathbb{AG}}$ *(all partial action profiles).* $\forall s, s' \in S \forall \delta_G : (s, s') \in \tau_{\delta_G}$ *iff* $\{\psi \mid [\delta_G]\psi \in s\} \cup \{Y\varphi \mid \varphi \in s\} \subseteq s'$,
- $L^c(p) = \{s \in S^c \mid p \in s\}$.

In this pseudo model, transition- and probability functions are not defined. Instead, it contains a Kripke-style binary relation τ. More precisely, a set of binary relations $\{\tau_{\delta_G}\}$ for any δ_G. These relations indexed by (partial) action profiles provide alternative representation of the transition function. Now, let's show that this representation is correct.

Lemma 1 (Transition function). *τ represents a well-defined transition function.*

Proof. At first, we want to replace binary relation τ with a well-defined transition function o^c. But we need to define availability function act^c first.

Let $\forall s, \forall i : \delta_i \in act^c(i,s)$ if $\exists \delta$, s.t. $\delta|_i = \delta_i$ and $\tau_\delta(s) \neq \emptyset$. For each (i,s), $act^c(i,s)$ is non-empty by (AA1-AA2) and the consistency of s. So, act^c is well-defined.

Now, for any $s \in \overline{S^c_f}$ and any complete action profile δ available to $\mathbb{AG}$ in s according to act^c, let $o^c(s,\delta) = \tau_\delta(s)$. Since $s \in \overline{S^c_f}$, $\neg final \in s$. By (AA1) and the consistency of s, there exists at least one δ for which $\tau_\delta(s)$ is non-empty. By (AA3), for any combination δ'_H of available individual actions ($\{\delta_i = \delta'_H|_i \mid i \in H \;\&\; \tau_{\delta_i}(s) \neq \emptyset\}$), $\tau_{\delta'_H}(s)$ is non-empty. This property is also known as Independence of Choice. Finally, (AA4) enforces that $\tau_{\delta|_{G \cup H}}(s) \subseteq \tau_{\delta|_G}(s)$ for $G \cap H = \emptyset$. As a result, any $o^c(s, \delta_G)$ is a union of $o^c(s, (\delta_G \cup \delta_{(\mathbb{AG}-G)})$. Recall that o^c is a non-deterministic function, so $o^c(s,\delta)$ is not necessarily a singleton. Thus, we obtained availability and transition functions (act^c, o^c).

Note that each $s \in S^c$ has at most one predecessor. Recall that o^- denotes an inverse of o^c. Let φ_s and $\varphi_{s'}$ be characteristic formulae of s and some $s' \in o^-(s)$ respectively. Then $\varphi_{s'} \in s'$ and $\langle\delta\rangle\varphi_s \in s'$ for some δ. By (YA2) and (MP) $[\delta]Y\varphi_{s'} \in s'$. Then, by the construction of s and the fact that $s \in o(s', \delta)$, $Y\varphi_{s'} \in s$. Assume that there exists $s'' \neq s'$, such that $s'' \in o^-(s)$. Then, $Y\varphi_{s''} \in s$, which implies $Y\neg\varphi_{s'}$. This contradicts (YA1) axiom and violates the consistency of s. Thus, each $s \in S^c$ has at most one predecessor. □

The last missing component in the pseudo-model is a probability function P^c. Let

$$\tilde{\alpha}_s^{\psi} = \max\{\alpha : Pr(\psi) \geq \alpha \in s\} \text{ and } \tilde{\beta}_s^{\psi} = \min\{\beta : Pr(\psi) \leq \beta \in s\}$$

It is straightforward to see that $\forall \gamma \in A(\varphi), \gamma \leq \tilde{\alpha}_s^{\psi}$ implies $Pr(\psi) \geq \gamma \in s$ and $\gamma \geq \tilde{\beta}_s^{\psi}$ implies $Pr(\psi) \leq \gamma \in s$. It remains to prove the following lemma:

Lemma 2. *For all $s \in S^c$ and all $\psi \in cl(\varphi)$: $\tilde{\alpha}_s^{\psi} \in F$ and $\tilde{\alpha}_s^{\psi} = \tilde{\beta}_s^{\psi}$*

Proof. Assume $\tilde{\alpha}_s^{\psi} \notin F$. Then $r_i < \tilde{\alpha}_s^{\psi} < r_{i+1}$, for some $r_i, r_{i+1} \in F$. Then $Pr(\psi) > r_i \in s$ since $Pr(\psi) \geq \tilde{\alpha}_s^{\psi}$ implies $Pr(\psi) > r_i$. By (PA6) axiom and the consistency of s, $Pr(\psi) > r_i \in s$ implies $Pr(\psi) \geq r_{i+1} \in s$ which contradicts the definition of $\tilde{\alpha}_s^{\psi}$ (where $\tilde{\alpha}_s^{\psi} < r_{i+1}$), and thus $\tilde{\alpha}_s^{\psi} \in F$.

Now, assume that $\exists s \in S^c$ and $\exists \psi \in cl(\varphi) : \tilde{\alpha}_s^{\psi} < \tilde{\beta}_s^{\psi}$. Then $Pr(\psi) \geq \tilde{\alpha}_s^{\psi} \in s$ and $Pr(\psi) \leq \tilde{\alpha}_s^{\psi} \notin s$. The latter implies that $Pr(\psi) > \tilde{\alpha}_s^{\psi} \in s$ by maximality of s. If $\tilde{\alpha}_s^{\psi} = r_i \in F$, then $Pr(\psi) \geq r_i \in s$ and $Pr(\psi) > r_i \in s$. Then, by axiom (PA6), we know that $Pr(\psi) \geq r_{i+1} \in s$, and $r_{i+1} > \tilde{\alpha}_s^{\psi}$. This contradicts the definition of $\tilde{\alpha}_s^{\psi}$. Thus, $\tilde{\alpha}_s^{\psi} = \tilde{\beta}_s^{\psi}$ □

Now for any state $s \in S^c$ and any subset $X \subseteq S^c$, where $X = \|\psi\|$ we define a probability function P^c as $P^c(s)(X) = \tilde{\alpha}_s^{\psi}$. This function is defined everywhere since $\{\tilde{\alpha}_s^{\psi}\}$ and $\{\tilde{\beta}_s^{\psi}\}$ are defined for all $\psi \in cl(\varphi)$ and any $X \subseteq S^c$ is an equivalence class of some $\psi \in cl(\varphi)$. It remains to show that $P^c(s)$ is indeed a probability measure.

Lemma 3. *$P^c(s)$ is a probability measure, i.e. satisfies P1-P3 in Definition 1.*

Proof. Note that $Pr(\top) \geq 1$ is a theorem of our logic (by Nec_2) and since $S^c = \|\top\|$, it follows immediately that $P^c(s)(S^c) = 1$. So, P1 is satisfied. For finite additivity (P3), consider two disjoint subsets $X \cap Y = \emptyset$ of S^c and let $\|\psi_X\| = X$ and $\|\psi_Y\| = Y$. So, $\|\psi_X\| \cap \|\psi_Y\| = \emptyset$ and then $\|\psi_X\| \subseteq \|\neg\psi_Y\|$. Since $\|\varphi_1\| \subseteq \|\varphi_2\|$ iff $\vdash \varphi_1 \rightarrow \varphi_2$, we have $\vdash \psi_X \rightarrow \neg\psi_Y$, which is equivalent to $\vdash \neg(\psi_X \wedge \psi_Y)$. By ($\text{Nec}_2$) we have $\vdash Pr(\neg(\psi_X \wedge \psi_Y)) \geq 1$. Then, by (PA5) and (MP) we get

$$\vdash (Pr(\psi_X) > \alpha \wedge Pr(\psi_Y) \geq \beta) \rightarrow Pr(\psi_X \vee \psi_Y) > (\alpha + \beta).$$

Then

$$s \vdash (Pr(\psi_X) > \alpha \wedge Pr(\psi_Y) \geq \beta) \rightarrow Pr(\psi_X \vee \psi_Y) > (\alpha + \beta).$$

Let $d > 0$ be any positive rational number, such that $\forall r, r' \in F, |r - r'| > d$. Assume that $P^c(s)(X) = r_1$ and $P^c(s)(Y) = r_2$, for some $r_1, r_2 \in F$. Then $s \vdash Pr(\psi_X) \geq r_1$ and $s \vdash Pr(\psi_y) \geq r_2$, and thus $s \vdash Pr(\psi_X) > (r_1 - d)$ by (PA6). Then, we have

$$s \vdash Pr(\psi_X) > (r_1 - d) \wedge Pr(\psi_Y) \geq r_2, \text{ and by (MP):}$$

$$s \vdash Pr(\psi_X \vee \psi_Y) > (r_1 + r_2 - d).$$

From the latter, by the choice of d and (PA6) we have $Pr(\psi_X \vee \psi_Y) \geq (r_1 + r_2)$. So, for any s, and any $r_1, r_2 \in F$, if $P^c(s)(X) \geq r_1$ and $P^c(s)(Y) \geq r_2$, then $P^c(s)(X \cup Y) \geq (r_1 + r_2)$. The argument for another direction, is similar and uses (PA4) axiom instead of (PA5). Finally, P2 semantically follows from P1 and finite additivity. □

Now, *canonical model* M^c is a pseudo model $\mathfrak{M}$ endowed with availability and transition functions (act^c, o^c) defined in Lemma 1 and probability function P^c. It remains to demonstrate that M^c satisfies all remaining constraints, i.e. is a model of our logic.

Proposition 1. *P^c satisfies P4-P5 in Definition 1.*

Proof. Towards a contradiction, assume there exists $s \in S^c$ and its characteristics formula φ_s, s.t. $\tilde{\alpha}_s^{\varphi_s} = 0$. Then $Pr(\varphi_s) \leq 0 \in s$ or, equivalently, $Pr(\neg\varphi_s) \geq 1 \in s$. By ($\text{PA}_T$) axiom, $\neg\varphi_s \in s$. The latter violates the consistency of s, so our assumption is wrong and $\tilde{\alpha}_s^{\varphi_s} > 0$ for any s. Then, $P^c(s)(\{s\}) > 0$.

For P5, assume $P^c(s_1)(\varphi_{s_2}) > 0$ and $P^c(s_1) \neq P^c(s_2)$ for some $s_1, s_2 \in S^c$. Then, $Pr(\varphi_{s_2}) > 0 \in s_1$, and for some ψ, $Pr(\psi) = \alpha \in s_1$ and $Pr(\psi) \neq \alpha \in s_2$. By ($\text{PA}_{45}$) axioms and the consistency of s_1, $Pr(Pr(\psi) = \alpha) = 1$. But since $s_2 \notin \|Pr(\psi) = \alpha\|$, by additivity of P^c it holds that $P^c(s_1)(\{s_2\}) = 0$. This contradicts our initial assumption, and thus P^c satisfies P5. □

Proposition 2. *M^c satisfies C1 and C2 constraints.*

Proof. Assume (C1) does not hold, so $P(s_1)(\{s_2\}) = 0$ for some $s_1, s_2 \in o^c(s, \delta)$. Then, $Pr(\varphi_{s_2}) = 0 \in s_1$ and $\langle\delta\rangle Pr(\varphi_{s_2}) = 0 \in s$. By (AA5) and (MP), $[\delta]Pr(\varphi_{s_2}) = 0 \in s$. And since $s_2 \in o^c(s, \delta)$, $Pr(\varphi_{s_2}) = 0 \in s_2$. The latter contradicts P4, and thus P^c satisfies (C1).

For any $s \in S^c$ and δ, s.t. $o^c(s, \delta) \neq \emptyset$, consider ψ, s.t. $\|\psi\| = o^c(s, \delta)$. Then, $[\delta]\psi \in s$. By (AA6), $[\delta]Pr(\psi) = 1 \in s$, and $Pr(\psi) = 1 \in s'$ for any $s' \in \|\psi\|$. Then, $Pr(s')(\|\psi\|) = 1$, satisfying (C2). □

Proposition 3. *For all $i \in \mathbb{AG}$, $\sim_i^c$ is an equivalence relation.*

Proof. The reflexivity holds trivially, since $K_i\psi \in s$ iff $K_i\psi \in s$. For transitivity, let $s_1 \sim_i^c s_2$ and $s_2 \sim_i^c s_3$. Then, by the construction of $\sim_i^c$, $K_i\psi \in s_1$ iff $K_i\psi \in s_2$ and $K_i\psi \in s_2$ iff $K_i\psi \in s_3$. Then $K_i\psi \in s_1$ iff $K_i\psi \in s_3$ and then $s_1 \sim_i^c s_3$. To show that $\sim_i^c$ is symmetric, assume that $s_1 \sim_i^c s_2$, then $K_i\psi \in s_1$ iff $K_i\psi \in s_2$. Then, trivially $s_2 \sim_i^c s_1$. □

Lemma 4 (Truth lemma). *For all $\psi' \in cl(\varphi), s \in S^c$:*

$$(M^c, s) \Vdash \psi' \textit{ iff } \psi' \in s$$

Proof. The proof is organized by the induction on ψ'. Cases for $\psi' = p$ and boolean are trivial.

CASE $\psi' = Pr(\psi) \geq \alpha$. For right-to-left direction, $Pr(\psi) \geq \alpha \in s$ implies $P^c(s)(\|\psi\|) \geq \alpha$ by the construction. Then

$$(M^c, s) \Vdash Pr(\psi) \geq \alpha.$$

For the other direction, $\mathcal{M}^c, s \Vdash Pr(\psi) \geq \alpha \Leftrightarrow P^c(s)(\|\psi\|) \geq \alpha$. By the construction of P^c and $\tilde{\alpha}_s^{\psi}$ we know that $P^c(s)(\|\psi\|) = \tilde{\alpha}_s^{\psi}$ and $\tilde{\alpha}_s^{\psi} \geq \alpha$. Then, $Pr(\psi) \geq \alpha \in s$.

CASE $\psi' = [\delta_G]\psi$. Let $(M^c, s) \Vdash [\delta_G]\psi$. Then, for all $s' \in o^c(s, \delta_G)$, $(M^c, s') \Vdash \psi$. By the previous induction step, $\psi \in s'$. And since $s' \in o^c(s, \delta_G)$, it holds that $(M^c, s') \Vdash \mathrm{Y}\varphi$ iff $(M^c, s) \Vdash \varphi$. Again, by previous induction step, $\mathrm{Y}\varphi \in s'$ for any $\varphi \in s$. Then,

$$\{\psi' \mid [\delta_G]\psi' \in s\} \cup \{\mathrm{Y}\varphi \mid \varphi \in s\} \subseteq s',$$

and by the construction of o^c, $[\delta_G]\psi \in s$.

For the other direction, assume $[\delta_G]\psi \in s$. Consider a set of formulas $Suc = \{\psi' \mid [\delta_G]\psi' \in s\}$. Note that by (YA2) for any $\varphi \in s$, $[\delta_G]\mathrm{Y}\varphi \in s$. So, $\mathrm{Y}\varphi \in Suc$. By the construction of the canonical

model, for any $s' \in S^c$, such that $s' \in o^c(s, \delta_G)$, it holds that $Suc \subseteq s'$. And since $\psi \in Succ$, $\psi \in s'$. By the previous induction step, $(M^c, s') \Vdash \psi$. Thus, $(M^c, s) \Vdash [\delta_G]\psi$.

CASE $\psi' = \mathsf{Y}\psi$. By contraposition, assume $(M^c, s) \nVdash \mathsf{Y}\psi$. Then, for all $s^- \in o^-(s)$, $(M^c, s^-) \nVdash \psi$. By the previous induction step, $\psi \notin s^-$, and then $\neg\psi \in s^-$. By the construction of o^c, $s^- \in o^-(s)$ implies that $\{\mathsf{Y}\varphi \mid \varphi \in s^-\} \subseteq s$. Then, $\mathsf{Y}\neg\psi \in s$. By (YA1), $\neg\mathsf{Y}\psi \in s$. Finally, by the consistency of s, $\mathsf{Y}\psi \notin s$.

For the other direction, let $(M^c, s) \Vdash \mathsf{Y}\psi$. So, $\exists s^- \in S^c$, s.t. $s^- \in o^-(s)$ and $(M^c, s^-) \Vdash \psi$. Then, by the construction of o^c,

$$\{\varphi \mid [\delta_G]\varphi \in s^-\} \cup \{\mathsf{Y}\varphi \mid \varphi \in s^-\} \subseteq s.$$

By the previous induction step, $\psi \in s^-$, and then $\mathsf{Y}\psi \in s$.

CASE $\psi' = K_i\psi$. $K_i\psi \in s$ iff $\forall s' : s \sim_i^c s' \Rightarrow K_i\psi \in s'$ by the construction. $K_i\psi \in s'$ implies $\psi \in s'$ by (T) axiom and the consistency of s'. By previous induction step, $\psi \in s'$ iff $(M^c, s') \Vdash \psi$. Thus, $K_i\psi \in s$ iff $\forall s' : s \sim_i^c s' \Rightarrow M^c, s' \Vdash \psi$. And hence $K_i\psi \in s$ iff $(M^c, s) \Vdash K_i\psi$. □

Theorem 1 (Completeness). L *is complete:* $\Vdash \varphi$ *iff* $\vdash \varphi$.

Proof. For any formula $\nvdash \varphi$, construct a model M' for $\neg\varphi$. From the Truth Lemma it follows that $M', s \Vdash \neg\varphi$ for some $s \in S$. Then $\nVdash \varphi$. The other direction follows from the soundness of (AX). □

Thus, any valid formula is derivable in L. Note also that from the proof of Theorem 1 it follows that the satisfiability problem for L is decidable. Formally, the satisfiability problem is: given $\varphi \in \mathsf{L}$, decide whether there exists (M, s), such that $(M, s) \Vdash \varphi$. We know that any satisfiable formula φ is satisfiable in a model of size not greater than the size of M^c from the proof of Theorem 1. Note that $|M^c| = O(2^{|cl(\varphi)|})$ and $|cl(\varphi)| = O(2^{|A(\varphi)|})$. And since there are only finitely many models M' of size $|M'| \leq 2^{2^{|A(\varphi)|}}$, the following result follows immediately:

Proposition 4. *Satisfiability problem for* L *is decidable.*

It is also easy to see that the model-checking problem for our logic is solvable in polynomial time. Verifying $Pr(\psi) \geq \alpha$ takes $O(|M|)$; $[\delta_G]\psi$ requires checking all outgoing transitions for δ_G and thus requires $O(|M|)$ steps; cases for $\mathsf{Y}\psi$ and $\mathsf{K}_i\psi$ are also solvable in $O(|M|)$. Thus, given (M, s) and φ, checking $(M, s) \Vdash \varphi$ requires polynomial time in $O(|M| \cdot |\varphi|)$.

6 CONCLUSION AND DISCUSSION

In this paper we proposed several notions of multi-agent responsibility for taking risks by contributing to an unnecessarily high chance of occurrence of some harmful event, even if the event does not necessarily happen. We proposed a logical framework in which all those notions can be represented, provided a complete axiomatization for the logic, and showed that it has a decidable satisfiability problem and an efficient model-checking procedure. Our approach is similar to [14], which extends Coalition logic [29] and in which agents are held responsible if they do not keep the risk of occurrence of of some undesirable event below a specified numerical threshold. We have shown that this notion can also be represented in our framework. Recently, other notions of responsibility that do not require an undesirable outcome to be necessarily realized have been proposed. For example, the notions of weak and strong passive responsibility based on dominant and best-effort strategies were proposed in [6]. These notions evaluate whether the agent did her best for avoiding an undesirable outcome, but do not take into account whether the outcome is actually realized. In contrast to our framework, this approach deals with deterministic outcomes and single-agent scenarios only. Another approach to responsibility in probabilistic setting has been proposed in [22]. Their framework is based on Probabilistic ATL logic and allows to encode notions of a degree of active and passive responsibility in the sense of [28]. Unlike our notions of passive responsibility, where a group of agents is responsible for the outcome if they could prevent the outcome independently of what other agents do, [22] assumes that the agent is passively responsible if, keeping all other agents' actions fixed, the agent could avoid the outcome. So, their degree of responsibility is the ratio of probability of the fixed actions of other agents violating the outcome relative to the probability of all possible behaviours violating the outcome. While both our approach and [22] compare probabilities of outcomes for different actions, [22] is focused on STIT-like rather than strategic setting.

We believe the restriction to a fixed finite set F of probabilities is reasonable in many scenarios where risk of a harmful outcome is assessed by a finite number of qualitative categories (e.g. low, medium, high). For example, side-effects of medicines list "fewer than 1 in 1000", "fewer than 1 in 10000", etc. We did not introduce that constraint because of issues in axiomatizing our language, but because of challenges in representing our proposed notions of responsibility by formulas of the language. Indeed, the counterfactual nature of these notions requires comparison of probabilities of a formula (describing a harmful event) in *different states*. To the best of our knowledge, virtually all probability logics do not have that ability. That holds even for the languages with the qualitative probability operator [16, 27, 31], or with comparison of linear combinations of probabilities [9] – they can compare probabilities of different formulas in the same state, but they are unable to compare probabilities of the same formula in different states. The only notable exception is the work of Halpern and Pucella [18], which allows relating probabilities in two consecutive time instants, which is obtained by extending syntax with polynomial probability terms and first-order quantification over probability values. Using those extensions, we can, for example, represent responsibility for not choosing the safest option that minimizes chance of occurrence of a harmful event φ as $\exists x(Pr(\varphi) \geq x \wedge \mathsf{Y}(\exists_{\delta_G} E_G[\delta_G]Pr(\varphi) < x \wedge \forall_{H \subset G} \neg\exists_{\delta_H} E_H[\delta_H]Pr(\varphi) < x))$.

Instead of such a heavy extension of the language, whose axiomatization would require all valid formulas about real closed fields [18], we opted to stay at the level of purely propositional logic with restriction to a fixed finite set F of probabilities. With that restriction, we were able to replace $\exists x$ in the above formula with a finite disjunction over values from F, obtaining our formula $Resp_G^{min}$.

Finally, although our novel notions of responsibility have zero tolerance for increasing risk, our logic can also be used to model alternative definitions with various tolerance levels. For example, the formula $\exists_{r \in F}(Pr(\varphi) \geq r + \epsilon \wedge \mathsf{Y}(\exists_{\delta_G} E_G[\delta_G]Pr(\varphi) < r \wedge \forall_{H \subset G} \neg\exists_{\delta_H} E_H[\delta_H]Pr(\varphi) < r))$ has the same intuition as $Resp_G^{min}$, with the exception that it allows the increase of risk for at most a tolerance level ϵ.

REFERENCES

[1] 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* (São Paulo, Brazil) *(AAMAS '17)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1091–1099.

[2] Christel Baier, Florian Funke, and Rupak Majumdar. 2021. A Game-Theoretic Account of Responsibility Allocation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Zhi-Hua Zhou (Ed.). International Joint Conferences on Artificial Intelligence Organization, 1773–1779. https://doi.org/10.24963/ijcai.2021/244 Main Track.

[3] Nils Bulling and Mehdi Dastani. 2013. Coalitional Responsibility in Strategic Settings. In *Computational Logic in Multi-Agent Systems*, João Leite, Tran Cao Son, Paolo Torroni, Leon van der Torre, and Stefan Woltran (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 172–189.

[4] Hana Chockler and Joseph Y. Halpern. 2004. Responsibility and Blame: A Structural-Model Approach. *Journal of Artificial Intelligence Research* 20 (2004), 93–115.

[5] Mehdi Dastani and Vahid Yazdanpanah. 2023. Responsibility of AI Systems. *AI & SOCIETY* 38, 2 (01 Apr 2023), 843–852. https://doi.org/10.1007/s00146-022-01481-4

[6] Giuseppe De Giacomo, Emiliano Lorini, Timothy Parker, and Gianmarco Parretti. 2025. Responsibility Anticipation and Attribution in LTLf. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, James Kwok (Ed.). International Joint Conferences on Artificial Intelligence Organization, 47–55. https://doi.org/10.24963/ijcai.2025/6 Main Track.

[7] Tiago de Lima, Lambér Royakkers, and Frank Dignum. 2010. A logic for reasoning about responsibility. *Logic Journal of the IGPL* 18, 1 (01 2010), 99–117. https://doi.org/10.1093/jigpal/jzp073

[8] Virginia Dignum. 2019. *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Springer Nature.

[9] Ronald Fagin and Joseph Y. Halpern. 1994. Reasoning about knowledge and probability. *J. ACM* 41, 2 (March 1994), 340–367. https://doi.org/10.1145/174652.174658

[10] Ronald Fagin, Joseph Y. Halpern, and Nimrod Megiddo. 1990. A logic for reasoning about probabilities. *Information and Computation* 87, 1 (1990), 78–128. https://doi.org/10.1016/0890-5401(90)90060-U Special Issue: Selections from 1988 IEEE Symposium on Logic in Computer Science.

[11] Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Vardi. 1995. *Reasoning about Knowledge*. MIT Press, Cambridge, Massachusetts.

[12] M. Fattorosi-Barnaba and G. Amati. 1987. Modal operators with probabilistic interpretations. *Studia Logica* 46 (1987), 383–393. https://doi.org/10.1007/BF00370648

[13] Michael J. Fischer and Richard E. Ladner. 1977. Propositional Modal Logic of Programs. In *Proceedings of the 9th STOC*. ACM, 286–294. https://doi.org/10.1145/800105.803418

[14] Maksim Gladyshev, Natasha Alechina, Mehdi Dastani, and Dragan Doder. 2023. Group Responsibility for Exceeding Risk Threshold. In *Proceedings of the 20th International Conference on Principles of Knowledge Representation and Reasoning*. 322–332. https://doi.org/10.24963/kr.2023/32

[15] Maksim Gladyshev, Natasha Alechina, Mehdi Dastani, and Dragan Doder. 2025. Reasoning about group responsibility for exceeding risk threshold in one-shot games. *Inf. Comput.* 303 (2025), 105257. https://doi.org/10.1016/J.IC.2024.105257

[16] Peter Gärdenfors. 1975. Qualitative Probability as an Intensional Logic. *Journal of Philosophical Logic* 4, 2 (1975), 171–185. http://www.jstor.org/stable/30226116

[17] Joseph Y. Halpern and Max Kleiman-Weiner. 2018. Towards Formal Definitions of Blameworthiness, Intention, and Moral Responsibility. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence* (São Paulo, Brazil) *(AAAI-18)*. 1853–1860.

[18] Joseph Y. Halpern and Riccardo Pucella. 2006. A Logic for Reasoning about Evidence. *J. Artif. Intell. Res.* 26 (2006), 1–34.

[19] Aviad Heifetz and Philippe Mongin. 2001. Probability Logic for Type Spaces. *Games and Economic Behavior* 35, 1 (2001), 31–53. https://doi.org/10.1006/game.1999.0788

[20] Luc Lismont and Philippe Mongin. 1994. A non-minimal but very weak axiomatization of common belief. *Artificial Intelligence* 70, 1 (1994), 363–374. https://doi.org/10.1016/0004-3702(94)90111-2

[21] Emiliano Lorini, Dominique Longin, and Eunate Mayor. 2013. A logical analysis of responsibility attribution: emotions, individuals and collectives. *Journal of Logic and Computation* 24, 6 (12 2013), 1313–1339. https://doi.org/10.1093/logcom/ext072

[22] Chunyan Mu, Muhammad Najib, and Nir Oren. 2025. Responsibility-aware strategic reasoning in probabilistic multi-agent systems. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'25/IAAI'25/EAAI'25)*. AAAI Press, Article 2594, 9 pages. https://doi.org/10.1609/aaai.v39i22.34492

[23] Pradeep K. Murukannaiah, Nirav Ajmeri, Catholijn M. Jonker, and Munindar P. Singh. 2020. New Foundations of Ethical Multiagent Systems. In *AAMAS*. International Foundation for Autonomous Agents and Multiagent Systems, 1706–1710.

[24] Pavel Naumov and Jia Tao. 2020. An epistemic logic of blameworthiness. *Artificial Intelligence* 283 (2020), 103269. https://doi.org/10.1016/j.artint.2020.103269

[25] Pavel Naumov and Jia Tao. 2021. Two Forms of Responsibility in Strategic Games. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Zhi-Hua Zhou (Ed.). International Joint Conferences on Artificial Intelligence Organization, 1989–1995. https://doi.org/10.24963/ijcai.2021/274 Main Track.

[26] Pavel Naumov and Jia Tao. 2023. Counterfactual and seeing-to-it responsibilities in strategic games. *Annals of Pure and Applied Logic* 174, 10 (2023), 103353. https://doi.org/10.1016/j.apal.2023.103353

[27] Zoran Ognjanovic, Aleksandar Perovic, and Miodrag Raskovic. 2008. Logics with the Qualitative Probability Operator. *Log. J. IGPL* 16, 2 (2008), 105–120.

[28] Timothy Parker, Umberto Grandi, and Emiliano Lorini. 2023. Anticipating Responsibility in Multiagent Planning. In *26th European Conference on Artificial Intelligence (ECAI 2023)*, Vol. 372. IOS Press.

[29] Marc Pauly. 2002. A Modal Logic for Coalitional Power in Games. *Journal of Logic and Computation* 12, 1 (02 2002), 149–166. https://doi.org/10.1093/logcom/12.1.149 arXiv:https://academic.oup.com/logcom/article-pdf/12/1/149/3657514/120149.pdf

[30] Lambèr Royakkers and Jesse Hughes. 2020. Blame it on me. *Journal of Philosophical Logic* 49, 2 (01 Apr 2020), 315–349. https://doi.org/10.1007/s10992-019-09519-7

[31] Krister Segerberg. 1971. Qualitative Probability in a Modal Setting. In *Proceedings of the Second Scandinavian Logic Symposium*, J.E. Fenstad (Ed.). Studies in Logic and the Foundations of Mathematics, Vol. 63. Elsevier, 341–352. https://doi.org/10.1016/S0049-237X(08)70852-8

[32] Qi Shi. 2024. Responsibility in Extensive Form Games. In *AAAI*. AAAI Press, 19920–19928.

[33] Helen Smith. 2020. Clinical AI: opacity, accountability, responsibility and liability. *AI & SOCIETY* (2020), 1–11.

[34] I. Sommerville, D. Cliff, R. Calinescu, J. Keen, T. Kelly, M. Kwiatkowska, J. Mcdermid, and R. Paige. 2012. Large-scale Complex IT Systems. *Communication of the ACM* 55, 7 (2012), 71–77.

[35] Wiebe van der Hoek. 1997. Some considerations on the logics P_FD A logic combining modality and probability. *Journal of Applied Non-Classical Logics* 7, 3 (1997), 287–307. https://doi.org/10.1080/11663081.1997.10510916

[36] Vahid Yazdanpanah and Mehdi Dastani. 2016. Distant Group Responsibility in Multi-agent Systems. In *PRIMA 2016: Principles and Practice of Multi-Agent Systems*, Matteo Baldoni, Amit K. Chopra, Tran Cao Son, Katsutoshi Hirayama, and Paolo Torroni (Eds.). Springer International Publishing, Cham, 261–278.

[37] Vahid Yazdanpanah, Mehdi Dastani, Wojciech Jamroga, Natasha Alechina, and Brian Logan. 2019. Strategic Responsibility Under Imperfect Information. In *Proceedings of the 18th International Conference on Autonomous Agents and Multi-Agent Systems* (Montreal QC, Canada) *(AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 592–600.

[38] Vahid Yazdanpanah, Enrico H. Gerding, Sebastian Stein, Mehdi Dastani, Catholijn M. Jonker, Timothy J. Norman, and Sarvapali D. Ramchurn. 2023. Reasoning about responsibility in autonomous systems: challenges and opportunities. *AI Soc.* 38, 4 (2023), 1453–1464.