Proceedings of the 20th International Conference on Principles of Knowledge Representation and Reasoning
Main Track

Group Responsibility for Exceeding Risk Threshold

Maksim Gladyshev , Natasha Alechina , Mehdi Dastani , Dragan Doder
Utrecht University, Utrecht, The Netherlands
{m.gladyshev, n.a.alechina, m.m.dastani, d.doder}@uu.nl

Abstract

The need for tools and techniques to formally analyze and trace the responsibility for unsafe outcomes to decision-making actors is urgent. Existing formal approaches assume that the unsafe outcomes for which actors can be held responsible are actually realized. This paper considers a broader notion of responsibility where unsafe outcomes are not necessarily realized, but their probabilities are unacceptably high. We present a logic combining strategic, probabilistic and temporal primitives designed to express concepts such as the risk of an undesirable outcome and being responsible for exceeding a risk threshold. We demonstrate that the proposed logic is complete and decidable.

1 Introduction

Safety of AI systems is a well-recognised and important concern. In multi-agent settings, where autonomous agents interact in complex ways, it is important to not only be able to determine whether an unsafe outcome is possible in principle, but also, when such an outcome occurs, determine why it occurred, which actions by which agents have caused it, and whether it could have been prevented. With the development of robots, self-driving cars and other autonomous agents, the urgency for formal tools and techniques to analyse the responsibility of AI systems has increased in last decades (Dignum 2019; Smith 2020; Dastani and Yazdanpanah 2023). This urgency has become clear in 2010 by the flash crash incident, where interacting high frequency algorithmic traders have led to extraordinary upheaval of U.S. equity markets (Sommerville et al. 2012). In this and other scenarios, the identification of agents, their performed actions, and their abilities that could have prevented the realization of the outcome are essential in determining which agent or group of agents can be held responsible for the realized outcome.

The keystone of contemporary discussions about responsibility is so-called 'principle of alternate possibilities' (Frankfurt 1969). This principle proposed by H. Frankfurt states that an agent can be held responsible for an action only if the agent could have acted differently. Following this principle we assume that responsibility of an outcome can be attributed to an agent only if the agent could (had strategic ability to) prevent the outcome. At the same time, groups of agents can have much more strategic power than single agents. This can create situations when no single agent is able to prevent an (undesirable) outcome, but a group of agents has the required ability. Following some existing approaches to formally analyse group responsibility and blameworthiness (Bulling and Dastani 2013; Naumov and Tao 2020), we consider a group of agents G to be responsible for some outcome φ if (1) φ actually holds, (2) group G could have prevented φ, and (3) G is minimal, i.e. there is no smaller group (w.r.t. set inclusion) that satisfies condition (2). While first two conditions are natural interpretation of the Frankfurt's principle of alternate possibilities, the third one is specific for coalitional settings. This condition is argued to be necessary (Lindahl 1977; Belnap and Perloff 1993; Yazdanpanah et al. 2019; Naumov and Tao 2020), because if any coalition can achieve some outcome, then all its super-coalitions can either. Without condition (3), the grand coalition would always be responsible if at least one of its sub-coalitions is. Note also that this definition does not require uniqueness of G, so there can be multiple groups responsible for φ.

These approaches assume that the responsibility for an (unsafe or undesirable) outcome can be assigned to a group of agents if the (unsafe or undesirable) outcome actually holds (condition 1). However, unsafe outcomes are not necessarily the states of affairs where a bad event has actually happened, but also the states of affairs where the probability that the bad event happens is unacceptably high. Thus, in ascribing the responsibility of unsafe outcome to a set of agents, it is not sufficient to consider situations where the bad event has actually happened, for example when two autonomous vehicles collided. It is just as important to consider 'near misses' and 'risky' situations, that is, the situations where the bad event has not happened, but its probability is unacceptably high, for example, when autonomous cars are set to drive with high speeds and close distances to each other. Such scenarios are also considered in legal practices where someone is held responsible for changing the probability of a state of affairs. For example, the criminal law of many legal systems does not only respond to already occurred harms, but also to the so-called "endangerment offences" that create the danger of harm (ten Voorde 2014; Feinberg 1984). In criminal law, the danger of harm is described as the state where the chance of (remote) harms is unacceptably high. Examples of endangerment offences in

criminal law are Child Endangerment law (Sim 2019) and Driving Endangerment law (Cunningham 2008). In such cases, an individual is held responsible because 1) the individual has created a risky situation where the probability of (remote) harm is unacceptably high, and 2) the individual could act differently to prevent such risky situation. Leaving very young children alone at home or drinking while driving are examples of endangerment offences that create risky situations. The existing formal approaches to responsibility do not capture the notion of endangerment, or the responsibility for a probabilistic state of affairs in general.

In order to formally investigate this broad notion of responsibility, we model such scenarios by associating a probability measure over propositions to each possible state. This allows us to talk about propositions which are false in a state (e.g. proposition 'vehicle_collided' is false) but have a high probability (e.g. because the vehicles are set to drive fast and close to each other). To elaborate this idea on a simple example, consider a state in which the probability of propositions 'head' (head is up) and 'tail' (tail is up) of a coin is $1/2$, regardless of which side of the coin is actually up in that state. The coin in this state is considered as a fair coin. An agent can then act to change probabilities of these propositions by tampering with the (fair) coin such that in the resulting state the probability of proposition 'head' is 0.99 (and the probability of proposition 'tail' is 0.01). Note that despite the low probability the proposition 'tail' can be true in the resulting state. This and other examples show scenarios where actions by agents or groups of agents can increase the probability of a particular event without necessarily making this event true. In such scenarios, an agent or a group of agents can be held responsible for creating risky and unsafe outcomes if they increase the risk of undesirable events.

In this paper, we propose and investigate a logic combining coalition ability operator from (Pauly 2002) with a probabilistic operator from (Heifetz and Mongin 2001) that allow reasoning about probabilities and their changes. In this logic, we can express that the probability of an event is greater than a certain number. For example, we can express that the probability of an accident has become greater than 10%. Although the proposed logic can be used to analyse various aspects of AI systems that involve reasoning about risks and probabilistic uncertainty in general, in this paper we use this logic to model and analyse the notion of (group) responsibility for risk.

The paper is organized as follows. In Section 2 we introduce concurrent game structures endowed with probabilities and additional temporal relation. In Section 3 we propose and discuss the definition of group responsibility for taking risk with respect to the proposed models. In Section 4 we propose a logic GRR and demonstrate that the operator $Resp_G(\varphi, \alpha)$ meaning "a group G is responsible for making a risk of φ higher than α" is expressible in GRR. Then we provide a (weakly) complete Hilbert-style axiomatisation of GRR and prove its decidability. Finally, in Section 5 we briefly overview existing works in this field and in Section 6 we discuss the proposals for future work.

2 Preliminaries: Models

At first, we need to define Concurrent Game Structures.

Definition 1 (CGS, pointed)**.** *A concurrent game structure (CGS) is a tuple $\Gamma = (\mathbb{AG}, S, Act, d, o)$, comprising a nonempty finite set of all agents $\mathbb{AG} = \{1, \dots, k\}$, a nonempty finite set of states S, where $S_0 \subseteq S$ denotes the set of initial states, and a nonempty finite set of (atomic) actions Act. Function $d : \mathbb{AG} \times S \longrightarrow \mathcal{P}(Act)\backslash\{\varnothing\}$ defines nonempty sets of actions available to agents at each state, and o is a (deterministic) transition function that assigns the outcome state $s' = o(s, (\alpha_1, \dots, \alpha_k))$ to a state s and a tuple of actions $(\alpha_1, \dots, \alpha_k)$ with $\alpha_i \in d(i, s)$ and $1 \leq i \leq k$, that can be executed by $\mathbb{AG}$ in s. For α_G that is an action profile of a non-grand coalition $G \subset \mathbb{AG}$, $o(s, \alpha_G)$ is defined as the set containing all outcomes of α_G completed by actions of agents outside the coalition. A pointed CGS is given by (Γ, s), where Γ is a CGS and s is a state in it.*

Given a CGS Γ, a positional (memoryless) *strategy* for an agent $a \in \mathbb{AG}$ or a-strategy, is a function $str_a : S \longrightarrow d(a, S)$. Given a coalition $G = \{a_1, \dots, a_m\}$, a positional strategy $str_G = \langle str_{a_1}, \dots, str_{a_m} \rangle$ maps each state from S to a tuple of actions $\langle str_{a_1}(s), \dots, str_{a_m}(s) \rangle$.

A model $\mathcal{M} = (\mathbb{AG}, S, Act, d, o, \mathsf{Past}, P, V)$ of our logic is a CGS endowed with a temporal relation Past, a probability function P and a valuation function V. For a temporal relation $\mathsf{Past} \subseteq S \times S$ we use $s' \in \mathsf{Past}(s)$ to denote $s\mathsf{Past}s'$, i.e. s' is one-step reachable from s by Past relation. We require that $\forall s, x, y \in S$: if $x \in \mathsf{Past}(s)$ and $y \in \mathsf{Past}(s)$, then $x = y$ i.e., each state has at most one temporal predecessor. By this reason we can use $s' \in \mathsf{Past}(s)$ and $s' = \mathsf{Past}(s)$ interchangeably. We use this extension to ensure that given a state $s \in S$ we can always identify the unique previous state $s' = \mathsf{Past}(s)$. This assumption is important since verifying responsibility requires evaluating strategic power of the agents on the previous step. To guarantee that this temporal relation Past and a transition function o are aligned, we impose the following constraints:

R0 $\forall s \in S, s_0 \in S_0 : s_0 \neq o(s, str_{\mathbb{AG}})$ for any strategy $str_{\mathbb{AG}}$

R1 $\forall s, s' \in S : s' \in \mathsf{Past}(s)$, then $s \notin S_0$

R2 $s \notin S_0 \Rightarrow \exists s' \in S : s' \in \mathsf{Past}(s)$

R3 $s' \in \mathsf{Past}(s) \Rightarrow \exists str_{\mathbb{AG}}, s = o(s', str_{\mathbb{AG}})$

Intuitively, R0 states that the grand coalition cannot enforce an initial state. R1 means that initial states have no past. Property R2 means that non-initial states have a past. And R3 implies that if s' is the past of s, then the grand coalition must be able to move from s' to s. As a result of this semantic choice each initial state $s_0 \in S_0$ generates a tree of transitions: each state has a unique Past predecessor and a non-empty set of o-successors.

We also require that our model is also endowed with a probability function $P : S \mapsto (2^S \mapsto [0, 1])$ assigning each state with a probability measure on S. Every $P(s)$ must satisfy the following conditions for all $s \in S$:

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

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

P3 $P(s)$ is (finitely) additive, i.e. $P(s)(\bigcup_{0\le i\le m} X_i) = \sum_{0\le i\le m} P(s)(X_i)$, where $X_i \cap X_j = \varnothing$ for any $i \neq j$,

P4 $P(s)$ is reflexive, i.e. $P(s)(\{s\}) > 0$,

P5 $P(s)(\{s'\}) > 0$ implies $P(s) = P(s')$.

The first three conditions are standard properties of probability, and reflexivity (the actual state of affairs has a non-zero probability) is a natural property of probability measure associated with states. Condition P5 enforces the fact that given a state $s \in S$ and another state $s' \in S$, such that s assigns s' a non-negative probability (i.e. s' belongs to the support set of S), it holds that s and s' share the same counterfactual probabilistic information, so $P(s) = P(s')$. Finally, V is a standard valuation function $V : Prop \longrightarrow 2^S$.

3 Responsibility for Risk

To define the notion of group responsibility for taking risks with respect to our models, let us consider a simple example.

Example 1 (Drink-driving)**.** *After a party Alice is facing a choice to take a taxi and make a safe trip home, leaving her car at the bar, or to break the law and choose unsafe (above acceptable risk level) trip driving her car home. Assume that Alice decides to take a taxi, but even though it was a safer choice, the taxi gets into an accident.*

Let crashed be a proposition 'Alice gets into an accident'. Alice has two possible actions $Act_a = \{taxi, drive\}$. We assume that the probability of getting into an accident in a taxi is .01, while this probability increases to .05 in case of drunk driving (the numbers are arbitrary). The second agent in this scenario is the environment, which decides what the outcome of the action is (we denote the environment's actions by e_1 and e_2).

Let us fix the acceptable risk level $\alpha_{\mathsf{crashed}} = .02$. Only the action $taxi$ is under the acceptable risk level in this example. Note that due to P5, $P(s_1) = P(s_2)$ and $P(s'_1) = P(s'_2)$.

$P(s_1)(\mathsf{crashed}) = P(s_2)(\mathsf{crashed}) = .01$ $\qquad$ $P(s'_2)(\mathsf{crashed}) = P(s'_1)(\mathsf{crashed}) = .05$

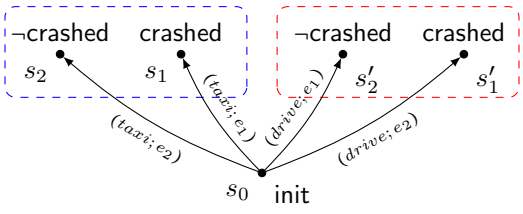

Figure 1: Model for Example 1. In this picture the temporal relation Past is omitted for readability: $s_0 = \mathsf{Past}(s')$ for all $s' \neq s_0$. Blue and Red rectangles are schematic representations of states assigned with a non-zero probability in s_1 and s'_1 respectively.

So, if Alice is in initial state s_0 and decides to take a taxi, then she may make a transition to s_1, in which the probability of an accident is low: $P(s_1)(\mathsf{crashed}) \le \alpha_{\mathsf{crashed}}$. But even though the risk is acceptable in s_1, the accident may happen. Alternatively, we can assume that the action $drive$ may lead to a state s'_2, in which the risk of crashed exceeds the acceptable risk level, but an accident does not happen.

Though in our example Alice may get home without an accident if she decided to drive, it can hardly be justified as a correct decision. In order to deal with such scenarios, when some (group of) agent can keep the probability of undesirable outcome sufficiently low, but does not do it, we propose the definition of group responsibility for taking risks.

Definition 2 (Responsibility for Risk)**.** *Let $s^- \in \mathsf{Past}(s)$. We say that a group G is responsible for exceeding acceptable risk $\alpha \in [0, 1]$ of a state of affairs $S' \subseteq S$ in s if the following conditions hold*

1. *$P(s, S') > \alpha$,*
2. *There is a strategy str_G for G, such that for all $s' \in o(s^-, str_G)$ it holds that $P(s', S') \le \alpha$,*
3. *G is minimal, i.e. no proper subset of G satisfies (2).*

We believe that this definition reflects the above mentioned intuitions naturally.

4 Logical Characterization

Now we are ready to introduce a logic for reasoning about group responsibility for taking risk (GRR). This logic is a fusion of Coalition Logic (Pauly 2002) and Probability Logic for Type Spaces (Heifetz and Mongin 2001) together with a temporal past operator. These ingredients are combined with the express purpose of being able to formalise Definition 2, that is, to be able to define the notion of being responsible for unacceptable level of risk.

4.1 Language and Semantics

Definition 3 (Language)**.** *The language of* GRR *is defined by the following grammar*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid L_\alpha\varphi \mid [G]\varphi \mid \boxminus\varphi,$$

where p ranges over $Prop \cup \{init\}$, G ranges over $2^{\mathbb{AG}}$ and α is any rational in $[0, 1]$.

We use this special proposition variable $init$ to distinguish initial states. $L_\alpha\varphi$ operator means "probability of φ is at least α". Derived operators $M_\alpha\varphi$ "probability of φ is at most α" and $I_\alpha\varphi$ "probability of φ is equal (identical) to α" can be defined as $M_\alpha\varphi \equiv L_{1-\alpha}\neg\varphi$ and $I_\alpha\varphi \equiv L_\alpha\varphi \wedge M_\alpha\varphi$ respectively. It also follows that $\neg L_\alpha\varphi$ and $\neg M_\alpha\varphi$ can be read as 'probability of φ is strictly smaller than α' and 'probability of φ is strictly greater than α' respectively. Formula $[G]\varphi$ reads as "group G can enforce φ to be true" and the formula $\boxminus\varphi$ reads as "φ was true at the previous step". The Boolean connectives $\vee, \rightarrow, \leftrightarrow, \bot$ and $\top$ are defined in the usual manner using $\neg$ and $\wedge$. The dual operator for $\boxminus$ is defined in the standard way: $\diamondminus\varphi \equiv \neg\boxminus\neg\varphi$.

Definition 4. *Given a model $\mathcal{M}$ and a state $s \in S$ we define $\vDash$ relation in the following way:*
$\mathcal{M}, s \vDash p$ iff $s \in V(p)$;
$\mathcal{M}, s \vDash \neg\varphi$ iff $\mathcal{M}, s \nvDash \varphi$;
$\mathcal{M}, s \vDash \varphi \wedge \psi$ iff $\mathcal{M}, s \vDash \varphi$ and $\mathcal{M}, s \vDash \psi$;
$\mathcal{M}, s \vDash [G]\varphi$ iff there is a strategy str_G for G, such that for

all $s' \in o(s, str_G)$ *it holds that* $\mathcal{M}, s' \vDash \varphi$*;*
$\mathcal{M}, s \vDash L_\alpha \varphi$ *iff* $P(s)([\varphi]^{\mathcal{M}}) \geq \alpha$*;*
$\mathcal{M}, s \vDash \boxminus \varphi$ *iff* $\forall s' \in \mathsf{Past}(s) : M, s' \vDash \varphi$.

We use $[\varphi]^{\mathcal{M}}$ as an abbreviation for $\{s \in S \mid \mathcal{M}, s \vDash \varphi\}$. We will omit superscript and write $[\varphi]$ if it is clear which model we are referring to.

4.2 Expressing Responsibility for Risk

This language allows us to define the new notion of responsibility for taking risk as a formula of GRR:

$$Resp_G(\varphi, \alpha) \equiv_{def} \neg M_\alpha \varphi \wedge \diamondminus [G] M_\alpha \varphi \wedge \bigwedge_{H \subset G} \diamondminus \neg [H] M_\alpha \varphi$$

We can check the correspondence between $Resp_G(\varphi, \alpha)$ and Definition 2 by checking the semantics of $\vDash$ relation from Definition 4.

1. $\mathcal{M}, s \vDash \neg M_\alpha \varphi$ iff $P(s)([\varphi]^{\mathcal{M}}) > \alpha$
2. $\mathcal{M}, s \vDash \diamondminus [G] M_\alpha \varphi$ iff in the unique $s^- \in \mathsf{Past}(s)$ there is a strategy str_G for G, such that for all $s' \in o(s^-, str_G)$ it holds that $P(s')([\varphi]^{\mathcal{M}}) \leq \alpha$
3. $\mathcal{M}, s \vDash \bigwedge_{H \subset G} \diamondminus \neg [H] M_\alpha \varphi$ iff no proper subset of G satisfies $[H] M_\alpha \varphi$ in the unique $s^- \in \mathsf{Past}(s)$

It can be easily seen that $Resp_G(\varphi, \alpha)$ operator does indeed encode Definition 2.

Proposition 1. *A group G is responsible for exceeding acceptable risk $\alpha \in [0, 1]$ of a state of affairs $[\varphi]^{\mathcal{M}} \subseteq S$ in s in a sense of Definition 2 if and only if $(\mathcal{M}, s) \vDash Resp_G(\varphi, \alpha)$.*

Returning to Example 1, we claim Alice to be responsible for not keeping the risk of crashed under $\alpha = .02$ in $(\mathcal{M}, s_2)$, because $\mathcal{M}, s_2 \vDash \neg M_{.02}\mathsf{crashed}$ and $\mathcal{M}, s_2 \vDash \diamondminus [\{a\}] M_{.02}\mathsf{crashed}$ and hence $\mathcal{M}, s_2 \vDash Resp_a(\mathsf{crashed}, .02)$. Even though an accident does not actually happen in s_2: $\mathcal{M}, s_2 \vDash \neg\mathsf{crashed}$. While for s_1 the situation is opposite: $\mathcal{M}, s_1 \vDash \mathsf{crashed}$ and $\mathcal{M}, s_1 \vDash \neg Resp_a(\mathsf{crashed}, .02)$.

4.3 Axiomatisation

Now we are ready to discuss the axiomatisation of GRR. As we already mentioned, GRR is essentially a fusion of coalition logic (Pauly 2002) and probability logic for type spaces (Heifetz and Mongin 2001). So the proof system for our logic combines original axioms for $[G]$ and L_α operators with fairly standard axioms for temporal operator $\boxminus$.

Axioms (CL1)-(CL5) are taken from (Pauly 2002), (CL6) guarantees that the grand coalition cannot enforce the initial state, $K_\boxminus$ and $U_\boxminus$ ensure that each state has a unique past, $1_\boxminus$ and $2_\boxminus$ say that initial states has no past and non-initial states have a past. $3_\boxminus$ says that the grad coalition can make a transition from past to the current state. (A1-A8) axioms are taken from (Heifetz and Mongin 2001). Note that the numeration of A1-A8 axioms may look odd, but we intentionally take it from the original paper. Finally, (T) ensures reflexivity while (4') and (5') (Fagin and Halpern 1994; Heifetz and Mongin 2001) guarantee that $P(s)$ satisfies condition (P5).

Axioms:	
(Taut)	All propositional tautologies
(CL1)	$\neg[G]\bot$
(CL2)	$[G]\top$
(CL3)	$\neg[\varnothing]\neg\varphi \to [\mathbb{AG}]\varphi$
(CL4)	$[G](\varphi \wedge \psi) \to [G]\varphi$
(CL5)	$[G_1]\varphi \wedge [G_2]\psi \to [G_1 \cup G_2](\varphi \wedge \psi)$, where $G_1 \cap G_2 = \varnothing$
(CL6)	$\neg[\mathbb{AG}]init$
($K_\boxminus$)	$\boxminus(\varphi \to \psi) \to (\boxminus\varphi \to \boxminus\psi)$
($U_\boxminus$)	$\diamondminus\varphi \to \boxminus\varphi$
($1_\boxminus$)	$init \to \boxminus\bot$
($2_\boxminus$)	$\neg init \to \diamondminus\top$
($3_\boxminus$)	$\neg init \wedge \varphi \to \boxminus[\mathbb{AG}]\varphi$
(A1)	$L_0\varphi$
(A2)	$L_\alpha\top$
(A5)	$L_\alpha\varphi \to \neg L_\beta\neg\varphi$, where $\alpha + \beta > 1$
(A8)	$\neg L_\alpha\varphi \to M_\alpha\varphi$
(T)	$L_1\varphi \to \varphi$
(4')	$L_\alpha\varphi \to L_1 L_\alpha\varphi$
(5')	$\neg L_\alpha\varphi \to L_1\neg L_\alpha\varphi$
Rules:	
(MP)	From φ and $\varphi \to \psi$, infer ψ
(Eq)	From $\varphi \leftrightarrow \psi$, infer $[G]\varphi \leftrightarrow [G]\psi$.
(Nec$_\boxminus$)	From φ, infer $\boxminus\varphi$
(A6)	From $\varphi \leftrightarrow \psi$ infer $L_\alpha\varphi \leftrightarrow L_\alpha\psi$
(B)	From $(\varphi_1, \ldots, \varphi_m) \leftrightarrow (\psi_1, \ldots, \psi_n)$, infer $\bigwedge_{i=1}^{m} L_{\alpha_i}\varphi_i \wedge \bigwedge_{j=2}^{n} M_{\beta_j}\psi_j \to L_\gamma\psi_1$, where $\gamma = (\alpha_1 + \cdots + \alpha_m) - (\beta_2 + \cdots + \beta_n)$

Table 1: The proof system for GRR.

The inference rule (B) of our probabilistic fragment requires additional clarification. The notation

$$(\varphi_1, \ldots, \varphi_m) \leftrightarrow (\psi_1, \ldots, \psi_n)$$

intuitively means that the same number of formulas from the set $\{\varphi_1, \ldots, \varphi_m\}$ is true as from the set $\{\psi_1, \ldots, \psi_n\}$. Formally, it is defined as follows.

By $\varphi^{(k)}$ we denote either the formula

$$\bigvee_{1 \leq l_1 < \cdots < l_k \leq m} (\varphi_{l_1} \wedge \cdots \wedge \varphi_{l_k})$$

or $\bot$ if $k > m$, similarly for $\psi^{(k)}$. Then

$$(\varphi_1, \ldots, \varphi_m) \leftrightarrow (\psi_1, \ldots, \psi_n)$$

stands for the formula

$$\bigwedge_{k=1}^{\max(m,n)} \varphi^{(k)} \leftrightarrow \psi^{(k)}.$$

In addition to the axioms and inference rules from Table 1 we will need to use some derived theorems to state the completeness result.

Proposition 2. *The following axiom and inference rule schemata can be derived from* GRR*:*
(A1+) $E_1\top$ *and* $\neg M_\alpha\top$*, where* $\alpha < 1$

(A2+) $E_0\bot$ *and* $\neg L_\alpha\bot$, *where* $\alpha > 0$
(A7) $L_\alpha\varphi \to L_\beta\varphi$, *where* $\alpha > \beta$
(A7+) $M_\alpha\varphi \to M_\beta\varphi$, *where* $\alpha < \beta$
(A12) $E_\alpha\varphi \to \neg E_\beta\varphi$, *where* $\alpha \neq \beta$
(B') *From* $((\varphi_1,\dots,\varphi_m) \leftrightarrow (\psi_1,\dots,\psi_n))$ *infer* $\left(\neg M_{\alpha_1}\varphi_1 \wedge \left(\bigwedge_{i=2}^{m} L_{\alpha_i}\varphi_i\right) \wedge \left(\bigwedge_{j=2}^{n} M_{\beta_j}\psi_j\right) \to \neg M_\gamma\psi_1\right)$, *where* $\gamma = (\alpha_1 + \dots + \alpha_m) - (\beta_2 + \dots + \beta_n)$.

The proof of this proposition can be found in (Heifetz and Mongin 2001).

4.4 Completeness and Decidability

For the completeness proof we need to consider a slightly different, but equivalent semantics. More precisely, we need to replace (Act, d, o) in $\mathcal{M}$ with an effectivity structure E. An effectivity structure as introduced in (Pauly 2002) assigns each state $s \in S$ with an effectivity function, i.e., $E : S \longrightarrow (2^{\mathbb{AG}} \longrightarrow 2^{2^S})$. This construction has a similar interpretation to the originally described transitions in our models, since effectivity functions describe choices available to each coalition G. Intuitively, in a state a coalition G is effective in achieving states $X \subseteq S$ iff this coalition has a joint strategy which will result in an outcome in X no matter what other agents do. More formally, given a coalition $G = \{a_1,\dots,a_m\}$ we say that G can enforce a set of states $S' \subseteq S$ in $s \in S$, i.e. $S' \in E(s)(G)$ if and only if there is a joint action $(\alpha_1,\dots,\alpha_m)$ for G, such that $o(s,(\alpha_1,\dots,\alpha_m)) \subseteq S'$.

So, the new (effectivity) model is a tuple $\mathcal{M}^e = (\mathbb{AG}, S, E, \mathsf{Past}, P, V)$. Let $\mathfrak{M}_e$ be the class of all such models. The interpretation of all operators in GRR except $[G]$ wrt $\mathcal{M}_e$ remains the same as in Definition 4. For $[G]\varphi$ we say that

$$\mathcal{M}, s \vDash [G]\varphi \text{ iff } [\varphi]^{\mathcal{M}} \in E(s)(G)$$

It was shown in (Goranko and Jamroga 2004) that transition- and effectivity-based semantics are equivalent for coalition logic. This implies straightforwardly that for any $\varphi \in \mathcal{L}(\mathsf{GRR})$, $\exists\mathcal{M} \in \mathfrak{M}, s \in S : \mathcal{M}, s \vDash \varphi$ iff $\exists\mathcal{M}' \in \mathfrak{M}_e, s' \in S' : \mathcal{M}', s' \vDash \varphi$.

To ensure this property we need to require effectivity functions E to be *truly playable*, which is a standard requirement for coalition logic CL (Pauly 2002; Goranko, Jamroga, and Turrini 2013).

Definition 5. *For any $s \in S$ the effectivity function $E(s)$ is truly playable iff it satisfies the following conditions:*

E1 $\forall G \subseteq \mathbb{AG} : \varnothing \notin E(s)(G)$ *(Liveness)*

E2 $\forall G \subseteq \mathbb{AG} : S \in E(s)(G)$ *(Safety)*

E3 $\overline{X} \notin E(s)(\varnothing)$ *implies* $X \in E(s)(\mathbb{AG})$, *where* $\overline{X}$ *denotes* $S \setminus X$ *(*$\mathbb{AG}$*-maximality)*

E4 $X \in E(s)(G)$ *and* $X \subseteq Y$ *implies* $Y \in E(s)(G)$ *(Outcome monotonicity)*

E5 *If* $G \cap D = \varnothing$, $X \in E(s)(G)$ *and* $Y \in E(s)(D)$, *then* $X \cap Y \in E(s)(G \cup D)$ *(Superadditivity)*

E6 $E^{nc}(s) \neq \varnothing$, *where* $E^{nc}(s)$ *is the non-monotonic core of the empty coalition:* $E^{nc}(s) = \{X \in E(s)(\varnothing) \mid \neg\exists Y(Y \in E(s)(\varnothing)$ *and* $Y \subsetneq X)\}$

We call an effectivity function *playable* if it only satisfies E1–E5. On finite domains E6 follows from E1 to E5 (Goranko, Jamroga, and Turrini 2013), so on finite domains an effectivity function is playable iff it is truly playable.

To prove completeness, we will show how to build a (finite) model for any formula φ such that $\neg\varphi$ is not derivable in GRR. This would imply that any valid formula is derivable. First, we fix a GRR-consistent formula φ and define a set $cl(\varphi)$ such that:

Definition 6 (Closure)**.** *For any* GRR*-consistent formula φ, $cl(\varphi)$ is the smallest set, such that*

- $cl(\varphi)$ *contains all subformulas of* φ*;*
- $cl(\varphi)$ *contains* $init, \boxminus\bot$*;*
- *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* α *appearing in* φ*. Then, for any formulas of the form* $L_\alpha\psi \in cl(\varphi)$ *and* $M_\alpha\psi \in cl(\varphi)$ *both* $L_\beta\psi \in cl(\varphi)$ *and* $M_\beta\psi \in cl(\varphi)$*, where* β *ranges over* $A(\varphi)$*;*
- $cl(\varphi)$ *is closed under single negation: if* $\psi \in cl(\varphi)$*, and* ψ *is not of the form* $\neg\xi$*, then* $\neg\psi \in cl(\varphi)$.

Let Ω be the set of all maximally consistent subsets s of $cl(\varphi)$ where each s is in addition extended by adding all formulas below that are consistent with s: for every set of consistent subsets $\{X_1,\dots,X_m\}$ of $cl(\varphi)$, add $\bigvee\{\bigwedge X_1 \cdots \bigwedge X_m\}$. Then, if $\neg init \in s$, and $\bigvee\{\bigwedge X_1 \cdots \bigwedge X_m\}$ is consistent with s, add $\boxminus[\mathbb{AG}]\bigvee\{\bigwedge X_1 \cdots \bigwedge X_m\}$ to s. Note that s remains consistent, becasue $\neg init \wedge \varphi \to \boxminus[\mathbb{AG}]\varphi$ is an axiom of GRR. Now s contains a conjunction of all the formulas originally in s, a disjunction defining all subsets of Ω it belongs to and a special formulas of the form $\boxminus[\mathbb{AG}]\varphi$ to ensure that property R3 holds. We denote a set of all formulas contained in Ω as $cl^*(\varphi) = \bigcup_{s\in\Omega} s$. So, Ω is a set of maximal consistent subsets of $cl(\varphi)$, such that each $s \in \Omega$ contains characteristic formulas of itself and all subsets of Ω it belongs to, together with formulas of the form $\boxminus[\mathbb{AG}]\varphi$. It it can be easily verified that $|\Omega| \leq 2^{O(|\varphi|\cdot|A(\varphi)|)}$.

Now, the following can be shown straightforwardly: (1) Ω is finite, (2) any subset of Ω is an equivalence class of some formula from $cl^*(\varphi)$ i.e., $\forall S' \subseteq \Omega\exists\psi \in cl^*(\varphi) : S' = [\psi]$, where $[\psi] = \{s \in \Omega \mid \psi \in s\}$ and (3) $[\varphi_1] \subseteq [\varphi_2]$ iff $\vdash \varphi_1 \to \varphi_2$.

We want to ensure that for any formula $\psi \in cl^*(\varphi)$ and any state $s \in \Omega$, s contains a well-defined probability interval for ψ. Formally, we want to require that for each $\psi \in cl^*(\varphi)$ and each $s \in \Omega$, s contains at least one formula of each forms $L_\alpha\psi$ and $M_\beta\psi$ (preserving consistency of s) such that $\max\{\alpha : L_\alpha\psi \in s\} \leq \min\{\beta : M_\beta\psi \in s\}$, where $\alpha, \beta \in A(\varphi)$. For any $\psi \in s$ there may be several choices of the formulas $L_\alpha\psi$ and $M_\alpha\psi$ satisfying the conditions, but it is sufficient to choose any of them. In particular, we can do it using the following simple algorithm. Fix a state $s \in \Omega$

and a formula $\psi \in cl^*(\varphi)$ and let $A_i(\varphi)$ be i's element of $A(\varphi)$. Starting from $i = 0$ we can run the following procedure: (1) if $s \cup L_{A_i(\varphi)}\psi$ is consistent, add $L_{A_i(\varphi)}\psi$ to s, (2) if $s \cup M_{1-A_i(\varphi)}\psi$ is consistent, add $M_{1-A_i(\varphi)}\psi$ to s, (3) $i = i + 1$. Repeating this procedure until $i > |A(\varphi)|$ will result in obtaining a desirable property mentioned above. Note that now some interval is defined for all $\psi \in cl^*(\varphi)$ such that $\psi \in s$ since if s is consistent, then $s \cup L_0\psi \cup M_1\psi$ is consistent as well since $L_0\psi$ and $M_1\psi$ are derivable in GRR. So, the first iteration of this procedure will always add $L_{A_i(\varphi)}\psi$ and $M_{1-A_i(\varphi)}\psi$ to s. In the next lemma we will show that $\max\{\alpha : L_\alpha\psi \in s\} - \min\{\beta : M_\beta\psi \in s\} \leq \frac{1}{q}$.

For any $s \in \Omega$ and any $\psi \in cl^*(\varphi)$ we define

$$\tilde{\alpha}^\psi = \max\{\alpha : L_a\psi \in s\} \text{ and } \tilde{\beta}^\psi = \min\{\beta : M_\beta\psi \in s\}$$

Now, we can state the following lemma:

Lemma 1. *1.* $\forall\gamma \in A(cl(\varphi)), \gamma \leq \tilde{\alpha}^\psi_s \Rightarrow L_\gamma\psi \in s$ *and* $\gamma \geq \tilde{\beta}^\psi_s \Rightarrow M_\gamma\psi \in s$

2. There are only two cases—either $\tilde{\alpha}^\psi_s = \tilde{\beta}^\psi_s$ *and* $I_{\tilde{\alpha}^\psi_s}\psi \in s$*, while* $I_\gamma\psi \notin s$ *for* $\gamma \neq \tilde{\alpha}^\psi_s$*, or* $\tilde{\alpha}^\psi_s < \tilde{\beta}^\psi_s$*, and* $I_\gamma\psi \notin s$*,* $\forall\gamma \in A(cl(\varphi))$

3. $\tilde{\beta}^\psi_s - \tilde{\alpha}^\psi_s \leq \frac{1}{q}$

Proof. The proof is essentially the same as the proof of Lemma A.2 from (Heifetz and Mongin 2001). (1) holds since $\gamma \leq \tilde{\alpha}^\psi_s \Rightarrow L_\gamma\psi \in s$ follows from (A7) and $\gamma \geq \tilde{\beta}^\psi_s \Rightarrow M_\gamma\psi \in s$ follows from (A7+). For (2) assume by contradiction that $\tilde{\alpha}^\psi_s > \tilde{\beta}^\psi_s$. Then from (1) we have that both $L_{\tilde{\beta}^\psi_s}\psi \in s$ and $M_{\tilde{\alpha}^\psi_s}\psi \in s$. Then both $E_{\tilde{\beta}^\psi_s}\psi \in s$ and $E_{\tilde{\alpha}^\psi_s}\psi \in s$ hold which is a contradiction by (A12). Thus for $\tilde{\alpha}^\psi_s = \tilde{\beta}^\psi_s$ it holds that $I_{\tilde{\alpha}^\psi_s}\psi \in s$ and for any $\gamma \neq \tilde{\alpha}^\psi_s$ $I_\gamma\psi \notin s$ by (A12). For the case where $\tilde{\alpha}^\psi_s < \tilde{\beta}^\psi_s$ the definition of $\tilde{\alpha}^\psi_s$ and $\tilde{\beta}^\psi_s$ implies that $I_\gamma\psi \notin s$ for any $\gamma \in A(cl(\varphi))$. For (3) suppose that $\tilde{\beta}^\psi_s - \tilde{\alpha}^\psi_s > \frac{1}{q}$. But then there is some $\gamma \in A(cl(\varphi)) \cap (\tilde{\alpha}^\psi_s, \tilde{\beta}^\psi_s)$. This implies that $\neg L_\gamma\psi \in s$ and $\neg M_\gamma\psi \in s$ contradicting (A8). □

Note that for any $s \in \Omega$ and any $\psi \in cl^*(\varphi)$, where ψ is of the form $L_\alpha\chi$ or $\neg L_\alpha\chi$ it holds that $\max\{\alpha : L_\alpha\psi \in s\} = 1$ by axioms 4' and 5' respectively. Now for any state $s \in \Omega$ and any subset $X \subseteq \Omega$, where $X = [\psi]$ we can define $\mathcal{F}^s_\psi$ to be either $\{\tilde{\alpha}^\psi_s\}$ if $\tilde{\alpha}^\psi_s = \tilde{\beta}^\psi_s$ or the open interval $(\tilde{\alpha}^\psi_s, \tilde{\beta}^\psi_s)$ if $\tilde{\alpha}^\psi_s < \tilde{\beta}^\psi_s$. Note that this construction is well-defined since $\{\tilde{\alpha}^\psi_s\}$ and $\{\tilde{\beta}^\psi_s\}$ are defined for all $\psi \in cl^*(\varphi)$ and any $X \subseteq \Omega$ is an equivalence class of some $\psi \in cl^*(\varphi)$. Let, for each set $s \in \Omega$, $P(s)$ be a probability measure on the subsets of Ω such that:

$$\forall X \subseteq \Omega, P(s)([\psi]) \in \mathcal{F}^s_\psi, \text{ where } X = [\psi] \qquad \text{(P)}$$

Lemma 2. *Probability measure* $P(s)$ *exists for all* $s \in \Omega$.

Proof. By the same technique as in (Heifetz and Mongin 2001). □

Now, we are ready to define a canonical model.

Definition 7 (Canonical Model)**.** *A (finite) canonical model is the tuple* $\mathcal{M}^c = (S^c, S^c_0, \mathbb{AG}^c, \mathsf{Past}^c, E^c, P^c, V^c)$*, where*

- $S^c = \Omega$, $S^c_0 = \{s \in S^c : init \in s\}$,
- $\mathbb{AG}^c = \mathbb{AG}$,
- $\forall s, s' \in S^c$: *choose one* s'*, s.t* $s' \in \mathsf{Past}^c(s)$ *from the set of* $\{s' \mid \forall \boxminus\psi \in cl(\varphi) : \boxminus\psi \in s \Rightarrow \psi \in s'\}$,
- *For* $s \in S^c, X \subseteq S^c$, $X \in E^c(s)(G) \Leftrightarrow \begin{cases}\exists\widetilde{\varphi} \subseteq X : s \vdash [G]\varphi & \text{for } G \neq \mathbb{AG}\\ \forall\widetilde{\varphi} \subseteq S^c\backslash X : s \nvdash [\varnothing]\varphi & \text{for } G = \mathbb{AG}\end{cases}$, *where* $\widetilde{\varphi} := \{s \in S^c \mid \varphi \in s\}$,
- *For* P^c *any probability function* P' *existing by Lemma 2,*
- $V^c(p) = \{s \in S^c \mid p \in s\}$.

Proposition 3. $P(s)$ *satisfies P4 and P5.*

Proof. To prove P4 it is sufficient to show that for all $s \in \Omega$ it holds that $\widetilde{\alpha}^{\varphi_s}_s > 0$, where φ_s is the characteristic formula of s. Assume that this is not the case. Then $\mathcal{F}^s_{\varphi_s} = \{\widetilde{\alpha}^{\varphi_s}_s\} = 0$. Then $M_0\varphi_s \in s$ which is equivalent (by definition) to $L_1\neg\varphi_s \in s$. Then, by T axiom it follows that $\neg\varphi_s \in s$ which is impossible due to consistency of s. Then our assumption was wrong and $P(s)$ is reflexive.

Condition P5 says that $P(s)$ is a Harsanyi type space (Heifetz and Mongin 2001) (this property is also called uniformity (Fagin and Halpern 1994)), for which probabilistic fragment of GRR together with (4') and (5') is known to be complete. For details see Theorem 5.2 in (Heifetz and Mongin 2001) or Theorem 4.2 in (Fagin and Halpern 1994). □

Proposition 4. $E^c(s)(G)$ *is truly playable for all* $s \in S^c$ *and all* $G \subseteq \mathbb{AG}$.

Proof. By the same technique as in (Pauly 2002). Note that our model $\mathcal{M}^c$ is finite, so it requires to check only E1-E5 conditions (Goranko, Jamroga, and Turrini 2013). □

Proposition 5. *For all* $s \in S^c$ *and all* $G \subseteq \mathbb{AG}$, $E^c(s)(G)$ *satisfies R0 and* R^c *satisfies R1-R3.*

Proof. (Sketch:) Property R0 is guaranteed by (CL6) and (CL3) axioms. R1 is enforced by $(1_\boxminus)$, R2 by $(2_\boxminus)$ and R3 by $(3_\boxminus)$. □

Proposition 3–Proposition 5 guarantee that $\mathcal{M}^c$ is a model of our logic. So, we are ready to establish the Truth lemma.

Lemma 3 (Truth Lemma)**.** *For all* $\psi \in cl(\varphi), M^c, s \vDash \psi$ *iff* $\psi \in s$.

Proof. The proof of the Truth Lemma is standard and will be done by induction of the formula size.

Case p trivial

Case booleans trivial

Case $[G]\varphi$. Consider the case when $G \neq \mathbb{AG}$. Assume that $\mathcal{M}^c, s \vDash [G]\varphi$. Then, by semantics there is $X \subseteq S^c : X \in E^c(s)(G)$ and for all $s' \in X : M^c, s' \vDash \varphi$. By construction of canonical model $\exists\widetilde{\varphi_0} \subseteq X : [G]\varphi_0 \in s$. By induction hypothesis $\{s' \mid M^c, s' \vDash \varphi\} = \widetilde{\varphi}$. Then $X \subseteq \widetilde{\varphi}$ and then $\widetilde{\varphi_0} \subseteq \widetilde{\varphi}$. Then $\vdash_{\mathsf{GRR}} \varphi_0 \to \varphi$. From Equivalence rule for $[G]$ operator, the fact that $[G]\varphi_0 \in s$ and consistency of s we derive that $[G]\varphi \in s$. Consider the other direction.

If $[G]\varphi \in s$, $\widetilde{\varphi} = \{s' : M^c, s' \vDash \varphi\}$ holds by induction hypothesis, then $\mathcal{M}^c, s \vDash [G]\varphi$ follows immediately.

For the case $G = \mathbb{AG}$, let $\mathcal{M}^c, s \vDash [G]\varphi$. Then, there is $X \subseteq S^c : X \in E^c(s)(G)$ and for all $s' \in X : M^c, s' \vDash \varphi$. Then $\forall \widetilde{\psi} \subseteq S^c \backslash X : [\varnothing]\psi \notin s$. And since $\widetilde{\neg\varphi} \subseteq S^c \backslash X$ it follows that $[\varnothing]\neg\varphi \notin s$ and then, by CL3 axiom, $\neg[G]\varphi \notin s$. By maximality of s, it follows that $[G]\varphi \in s$. For the other direction, let $[G]\varphi \in s$. Let $X = \{s \mid M^c, s \vDash \varphi\}$. We want to show that $\forall \widetilde{\psi} \subseteq S^c \backslash X, [\varnothing]\psi \notin s$. Note that $[\neg\varphi] = S^c \backslash X$ and then $\forall \widetilde{\psi} \subseteq S^c \backslash X, \vdash_{\mathsf{GRR}} \psi \to \neg\varphi$. Assume by contradiction that $\exists \widetilde{\psi} \subseteq S^c \backslash X : [\varnothing]\psi \in s$. But since $\psi \to \neg\varphi$, it must also hold that $[\varnothing]\neg\varphi \in s$ by (Eq). It contradicts our previous assumption that $[G]\varphi \in s$. Thus, $\forall \widetilde{\psi} \subseteq S^c \backslash X : [\varnothing]\psi \notin s$ and hence $\mathcal{M}^c, s \vDash [G]\varphi$.

Case $L_\alpha\varphi$. For right-to-left direction, $L_\alpha\varphi \in s$ implies $P^c(s)([\varphi]) \geq \alpha$ by the construction. Then $\mathcal{M}^c, s \vDash L_\alpha\varphi$. For the other direction, $\mathcal{M}^c, w \vDash L_\alpha\varphi \Leftrightarrow P^c(s)([\psi]) \geq \alpha$. By the construction of P^c and $\tilde{\alpha}^\psi_s$ we know that $P^c(s)([\psi]) = \tilde{\alpha}^\psi_s$ and $\tilde{\alpha}^\psi_s \geq \alpha$. Then, $L_\alpha\psi \in s$.

Case $\boxminus\varphi$. $\mathcal{M}^c, s \vDash \boxminus\varphi \Longleftrightarrow \forall s' \in \mathsf{Past}^c(s), M^c, s' \vDash \varphi$ by definition of $\vDash$ relation. $\forall s' \in \mathsf{Past}^c(s), M^c, s' \vDash \varphi \Longleftrightarrow \forall s' \in \mathsf{Past}^c(s), \varphi \in s'$ by previous induction step. $\forall s' \in \mathsf{Past}^c(s), \forall \boxminus\varphi \in cl(\varphi) : \boxminus\varphi \in s \Longleftrightarrow \varphi \in s'$. □

Theorem 1 (Completeness)**.** *Logic* GRR *is complete wrt* $\mathcal{M}^{\mathsf{GRR}}$*, i.e.* $\vDash \varphi$ *iff* $\vdash_{\mathsf{GRR}} \varphi$.

Proof. Right-to-left direction follows from the soundness of GRR. For left-to-right direction consider formula φ such that $\nvdash_{\mathsf{GRR}} \varphi$. Construct a model $\mathcal{M}$ for $\neg\varphi$. It follows from Lemma 3 that $\exists x \in S$, such that $\mathcal{M}, s \vDash \neg\varphi$. Then $\nvDash_{\mathcal{M}^{\mathsf{GRR}}} \varphi$ since $\mathcal{M} \in \mathcal{M}^{\mathsf{GRR}}$. □

Theorem 2 (Decidability)**.** *The satisfiability problem for* GRR *is decidable.*

Proof. The proof follows the technique presented in (Dautović, Doder, and Ognjanović 2021). We show that a formula φ is satisfiable iff it is satisfiable in one of finitely many 'solvable pre-structures' of a fixed size bounded by $|\varphi|$.

From the proof of Completeness theorem, we know that a formula φ is satisfiable iff it is satisfiable in a model $\mathcal{M} \in \mathcal{M}^e$ with at most $2^{|cl(\varphi)|}$ states. Thus, it is sufficient to check only models with $l \leq 2^{|cl(\varphi)|}$ states; however since our models include probability measures, there are infinitely many such models. In order to restrict the set of models to check to be finite, we will consider pre-structures which do not have probability measures, but where it is easy to check whether a corresponding measure does exist (in which case we call them solvable). The existence of one of such solvable pre-structures satisfying φ will guarantee the existence of a proper model (with a probability measure attached to each state) that satisfies φ.

Let $Prob(\varphi)$ be the set of all subformulas of φ of the form $L_\alpha\psi$ (here we assume that φ contains only the primitive probabilistic operators of the form L_α, since the other operators, like M_α, are introduced as abbreviations), and let $Prop(\phi)$ denote the set of propositional letters from φ. For every $l \leq 2^{|cl(\varphi)|}$ we consider pre-structures $\overline{M} = (\overline{S}, \overline{\mathbb{AG}}, \overline{\{\sim_i\}}_{i\in\mathbb{AG}}, \overline{R}, \overline{E}, pseudo - \overline{P}, \overline{V})$, where all items except $pseudo - \overline{P}$ and $\overline{V}$ are defined in a standard way. $\overline{V} : Prop(\varphi) \longrightarrow 2^{\overline{S}}$ is a valuation function restricted to $Prop(\varphi)$, and $pseudo - \overline{P}$ is 'emulating' a probability measure:
$pseudo - \overline{P} : \overline{S} \times Prob(\varphi) \longrightarrow \{true, false\}$.

It is clear that there are only finitely many pre-structures for each l. These pre-structures are not models of our logic, but we can check if a subformula of φ holds in some state s of this pre-structure using the $\vDash'$ relation which is defined in a standard way, except the case for $L_\alpha\psi$. For this case it is defined as follows:
$\overline{M}, s \vDash' L_\alpha\psi$ iff $pseudo - \overline{P}(s, L_\alpha\psi) = true$.

We will consider only those pre-stuctures $\overline{M}$ such that $\overline{M}, s \vDash' \varphi$ for some $s \in \overline{S}$. For each such $\overline{M}$ we want to check whether $\overline{M}$ can be extended to a model $\mathcal{M} = (\overline{S}, \overline{\mathbb{AG}}, \overline{\{\sim_i\}}_{i\in\mathbb{AG}}, \overline{R}, \overline{E}, P, V) \in \mathcal{M}$ of our logic.[1] In other words, we want to check if $pseudo - \overline{P}$ can be replaced by a probability function P such that P agrees with $pseudo - \overline{P}$, i.e., for every state s and every $\psi \in Prob(\varphi)$ we have $\mathcal{M}, s \vDash \psi$ iff $pseudo - \overline{P}(s, \psi) = true$. It is straightforward to check that for such $\mathcal{M}$ it holds that $\mathcal{M}, s \vDash \chi$ iff $\overline{M}, s \vDash' \chi$ whenever χ is a subformula of φ. For this purpose we consider special systems of linear equations and inequalities to define a probability measure $P(s)$ for each s:

(1) $P(s)(s_j) \geq 0$ for each $s_j \in \overline{S}$

(2) $\sum_{s_j \in \overline{S}} P(s)(s_j) = 1$

(3.1) $\sum_{s_j : \overline{M}, s_j \vDash' \psi} P(s)(s_j) \geq \alpha$ for every formula $L_\alpha\psi$ such that $pseudo - P(s, L_\alpha\psi) = true$

(3.2) $\sum_{s_j : \overline{M}, s_j \vDash' \psi} P(s)(s_j) < \alpha$ for every formula $L_\alpha\psi$ such that $pseudo - P(s, L_\alpha\psi) = false$

(4) $P(s)(s) > 0$

(5) for each $s' \in \overline{S}$, either (5.1) $P(s)(s') = 0$ or (5.2) for every $s_j \in \overline{S}$, $P(s)(s_j) = P(s')(s_j)$.

Inequality (1) ensures that the probability measure of any state is non-negative, and equation (2) guarantees that the sum of probability measures of all states is equal to 1. The inequalities (3.1) and (3.2) guarantee that P agrees with $pseudo - P$. For (3.1) and (3.2), we used the property that for every $X \subseteq \overline{S}$, $P(s)(X) = \sum_{s_j \in X} P(s)(s_j)$. Finally, (4) and (5) guarantee that every $P(s)$ satisfies P4 and P5.

Note that the equations and inequalities listed above form not one, but a number of finite systems of equations and inequalities. Indeed, adding (5) (for any two fixed s, s') to any system Sys results with a disjunction of two different extensions of Sys (one containing (5.1), and one containing (5.2)). For the purposes of our proof, it is sufficient to find at least one solution of one such system of linear equations and inequalities with the set of variables $\{P(s)(s') \mid s, s' \in \overline{S}\}$, and it is well known that the problem of solving systems of inequalities is decidable. So, given a pre-structure, we can check whether $pseudo - P$ corresponds to a system of inequalities that has a solution (is a solvable pre-structure).

[1] The way in which V extends $\overline{V}$ is obviously irrelevant.

It is straightforward to see that if there is a solvable pre-structure for φ, then φ is satisfiable.

The other direction: if φ is satisfiable, then there is a solvable pre-structure for φ with $2^{|cl(\varphi)|}$ states, is trivial since the canonical model for gives rise to a solvable pre-structure.

We have shown that φ is satisfiable iff there is a solvable pre-structure for φ with at most $l \leq 2^{|cl(\varphi)|}$ states. Since we have finitely many possibilities for the choice of l, and for every l there are only finitely many possibilities for the choice of pre-structure, we can check whether φ is satisfiable by examining all finitely many such solvable pre-structures. □

Finally, let the model checking problem be "given a $(\mathcal{M}, s, \varphi)$ check if $(\mathcal{M}, s) \vDash \varphi$". The following result can be established straightforwardly.

Proposition 6 (Model checking)**.** *The model checking problem for* GRR *is decidable in polynomial time.*

So, the problem of verifying whether a group G is responsible for the increased risk of φ in $(\mathcal{M}, s)$ is tractable.

5 Related Work

Several proposals to formalise such notions as responsibility or blameworthiness have been made in recent years. Chockler and Halpern (Chockler and Halpern 2004) studied the notions of a *degree* of responsibility and blame based on the definition of causality for the case of a single agent. Their definition of causality elaborates on the simple requirement of being able to prevent a state of affairs and replaces it with being able to prevent it *under some contingency*. For example, it is possible that agent a cannot prevent a state of affairs because even if a does not act to bring it about, another agent b would have done so. However an action by a can still be considered a cause of the state of affairs, under the contingency that b had chosen to act differently. The notion of responsibility is treated as a causal notion, which does not take into account the agent's epistemic state (whether the agent was aware of the consequences of his or her actions). The notion of blame takes the agent's epistemic state into account, by introducing a probability distribution over the models of the world according to the agent. If the agent assigns a high probability to the actual model of the world, then the degrees of responsibility and blame are very similar. However if an agent assigns a low probability to the actual state of affairs and actual effects of actions, then their degree of blame may be very low. In (Alechina, Halpern, and Logan 2017), these notions were applied to determining the degree of responsibility and blameworthiness in a multi-agent context. Halpern and Kleiman-Weiner (Halpern and Kleiman-Weiner 2018) studied the interplay of notions such as degree of blameworthiness and intention together with their connection to moral responsibility judgements. Logical modeling of interplay between knowledge and responsibility based on epistemic STIT logic (Herzig and Troquard 2006) was studied in (Ramírez Abarca and Broersen 2021) and (Lorini, Longin, and Mayor 2013). In strategic multi-agent settings, de Lima, Royakkers and Dignum (de Lima, Royakkers, and Dignum 2010) proposed a logical framework for reasoning about both forward- and backward-looking individual responsibility. In this framework the notions of responsibility were formalized as a combination of basic modalities such as agents' actions, abilities, obligations and knowledge. Later Royakkers and Hughes (Royakkers and Hughes 2020) extended this framework and formalized various notions of group responsibility and blameworthiness proposed by Van de Poel (van de Poel 2011). Yazdanpanah et al. (Yazdanpanah and Dastani 2016; Yazdanpanah et al. 2019) studied individual and group responsibility under imperfect information, and provided formal analyses for both forward and backward notions of responsibility. Naumov and Tao (Naumov and Tao 2020) provided a logical analysis of the interplay between blameworthiness and knowledge. They proposed a sound and complete logic for reasoning about blameworthiness in strategic games with imperfect information.

Our logic has a past operator, in common with many temporal logics. The operator is also somewhat similar to the converse modality in logics of actions, such as, for example, Propositional Dynamic Logic (Parikh 1978; Schild 1991; De Giacomo and Lenzerini 1994). Unlike the converse modality or converse action in (Deuser and Naumov 2021), our past is unique, in common with widely accepted intuitions in temporal logic: the past is linear, even if the future is branching. There is just one history that actually happened. In particular, if we are only talking about one step histories, there is a single state in the past (and not multiple alternative yesterdays). We stress that this choice is not forced by technical considerations but is deliberate and corresponds to the consensus in temporal logic about modelling the past.

On the technical side, this article belongs to a large body of work on combining modalities in order to be able to analyse AI systems. There are well-known logics that combine temporal and probabilistic modalities as well as temporal and strategic. Recently, logics combining strategic modalities with probabilities have also been proposed, but they concentrate on probabilities of the outcomes of actions and strategies, rather than strategies to enforce a particular probability distribution (Bulling and Jamroga 2009; Novák and Jamroga 2011; Huang, Su, and Zhang 2012; Naumov and Tao 2019; Aminof et al. 2019).

6 Discussion

In this paper we present a logic GRR for reasoning about group responsibility for taking risks. GRR demonstrates that such notion of responsibility can be formalized via a combination of primitive modalities such as strategic ability, time and probability rather than as a separate modal operator. We believe that the logic we propose in this paper is the first attempt to combine probabilistic logic in the sense of (Fagin and Halpern 1994) with strategic modalities. We provide a complete axiomatisation of the logic in order to facilitate the study of definable concepts and their properties. We also show that the logic has a decidable satisfiability problem, which makes it possible for the agents to reason automatically about whether they could be held responsible for the increased levels of risks. Finally, it holds that the model

checking problem for GRR is decidable in polynomial time. So, our framework can be combined with automated verification techniques and, thus, given a model it is always possible to verify if some group of agents is responsible in terms of the interpretation of responsibility discussed in our paper. Additional feature of our approach is the fact that we use only finitary languages and stay on the propositional level of reasoning avoiding first-order quantification in our syntax while working with probabilities. These properties are desirable for practical means in AI research and they are not always the case for probabilistic logics.

Here we mention a few properties of $Resp_G(\varphi, \alpha)$ operator. The first obvious observation suggests that if the group G is responsible for exceeding acceptable risk level α for φ, it is not always the case that G is responsible for exceeding a lower risk level α' for φ. This property obviously holds since 'being able to make a probability that φ holds at most α' does not imply 'being able to make a probability that φ holds at most α''. One can also expect that the opposite must be derivable in GRR, but surprisingly, it is not!

Proposition 7. $\not\models Resp_G(\varphi, \alpha) \rightarrow Resp_G(\varphi, \alpha')$*, where* $\alpha' > \alpha$.

At first glance this seems counterintuitive, because if some group G can enforce the probability of φ to be at most α, then it can automatically enforce it to be at most α', because $\vdash_{\mathsf{GRR}} [G]M_\alpha\varphi \rightarrow [G]M_{\alpha'}\varphi$ for $\alpha' > \alpha$. So, condition (2) of Definition 2 is satisfied. But the condition (1) can be violated: $P(s, S') > \alpha$ does not imply $P(s, S') > \alpha'$ for $\alpha' > \alpha$ and hence $\nvdash_{\mathsf{GRR}} \neg M_\alpha\varphi \rightarrow \neg M_{\alpha'}\varphi$. Next, consider how different groups interact. Assume that some group G is responsible for some risk level α of φ and another group D is responsible for another risk level β of φ. Then it is obviously not the case that together they are responsible for φ according to either α or β.

Proposition 8. $\models Resp_G(\varphi, \alpha) \wedge Resp_D(\varphi, \beta) \rightarrow \neg Resp_{G\cup D}(\varphi, \min(\alpha, \beta))$*, where* $D \cap G = \varnothing$.

Proposition 9. $\models Resp_G(\varphi, \alpha) \wedge Resp_D(\varphi, \beta) \rightarrow \neg Resp_{G\cup D}(\varphi, \max(\alpha, \beta))$*, where* $D \cap G = \varnothing$.

It is easy to see that both $Resp_{G\cup D}(\varphi, \min(\alpha, \beta))$ and $Resp_{G\cup D}(\varphi, \max(\alpha, \beta))$ violate minimality condition (3) from Definition 2, because antecedent guarantees that there are two proper subsets of $G \cup D$, namely G and D, that can enforce $P(s)([\varphi]) \leq \alpha$ and $P(s)([\varphi]) \leq \beta$ respectively. Finally, consider how $Resp_G(\varphi, \alpha)$ operator deals with Boolean connectives. In is natural to assume that if group G is responsible for (φ, α) and it is also responsible for $(\varphi \rightarrow \psi, \alpha)$, then G must be responsible for (ψ, α). Surprisingly, it is also not derivable in GRR!

Proposition 10. $\not\models Resp_G(\varphi, \alpha) \wedge Resp_G(\varphi \rightarrow \psi, \alpha) \rightarrow Resp_G(\psi, \alpha)$.

The idea of the proof is to construct a counterexample violating a minimality condition (3): there is no guarantee that G is a minimal coalition that could enforce $P(\psi) \leq \alpha$. So, there exists a model M and a state s, such that $M, s \models \Diamond[H]M_\alpha\psi$, where $H \subset G$. The last two propositions demonstrate that $Resp_G(\varphi, \alpha)$ operator behaves similarly for the case of conjunction.

Proposition 11. $\not\models Resp_G(\varphi, \alpha) \wedge Resp_G(\psi, \alpha) \rightarrow Resp_G(\varphi \wedge \psi, \alpha)$.

We can construct a counterexample as follows. Let $\alpha = 0$, so condition (2) holds if G can enforce $P(\varphi) = 0$. For simplicity, let $G = \{i\}$. Assume that i has three options: to enforce $P(\varphi) = 0$ and $P(\psi) = 1$, to enforce $P(\varphi) = 1$ and $P(\psi) = 0$, or to enforce $P(\varphi) = .5$ and $P(\psi) = .5$. If the last choice is made, then both $Resp_i(\varphi, \alpha)$ and $Resp_i(\psi, \alpha)$ hold, but not $Resp_i(\varphi \wedge \psi, \alpha)$.

Proposition 12. $\not\models Resp_G(\varphi \wedge \psi, \alpha) \rightarrow (Resp_G(\varphi, \alpha) \wedge Resp_G(\psi, \alpha))$.

For counterexample assume that $P(s)([\psi]) = 1$ in all states $s \in S$. And G can enforce $P(s)([\varphi]) = 0$. If the group ignores this choice and $\alpha = 0$, then $Resp_G(\varphi \wedge \psi, \alpha)$ holds since G could possibly enforce $P(s)([\varphi \wedge \psi]) = 0$, but $Resp_G(\psi, \alpha)$ does not hold, because G has no control over the probability of ψ.

These results demonstrate that it is often impossible to transfer responsibility to other groups of agents or events. In other words, if the group G is responsible for φ with a risk level α, then in most cases it does not imply any claim about responsibility of other group D for (φ, α) or any claim about responsibility of G for another state of affairs ψ and/or another risk level β.

Our work also has some limitations that inspire directions for future research. We build GRR over a Coalition logic CL which is essentially a Next-fragment of ATL logic (Alur, Henzinger, and Kupferman 2002). The choice of the use of Coalition logic was to consider strategic ability in the simplest abstract setting. The consequence of this is that responsibility is defined with respect to previous state, and not to an arbitrary state in the history. Other work, e.g., Yazdanpanah et al. (Yazdanpanah et al. 2019), have used ATL to define a group of agents responsible for an outcome if the group had an alternative to prevent the outcome at some point in the past. Another direction of future research is incorporating imperfect information in the spirit of (Jamroga 2003; Jamroga and van der Hoek 2004; Fervari et al. 2017; Naumov and Tao 2020). This however involves solving the problem of axiomatising CL or ATL under the strongly uniform strategies semantics.

In most probabilistic temporal logics and in Markov Decision Processes, probability distributions are over outcomes of non-deterministic actions. In this paper, we have chosen not to couple probability distributions to actions; our models admit free floating states that are not the outcome of any action by the agents but contribute to the probability distribution on its outcomes; or, the set of outcomes may be partitioned in two equivalence classes of probability distributions. We can however add an explicit environment agent as in Example 1 that is the source of non-determinism, and make the states in the outcome of each action of the grand coalition minus the environment agent to be exactly a single equivalence class where all states have the same probability distribution. This will make our semantics match the setting of the MDPs (where a transition by all agents is to a probability distribution over states). Axiomatisation of this setting is the topic of future work.

Acknowledgments

We thank the anonymous KR 2023 reviewers for their incisive and constructive comments.

References

Alechina, N.; Halpern, J. Y.; and Logan, B. 2017. Causality, responsibility and blame in team plans. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '17, 1091–1099. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Alur, R.; Henzinger, T. A.; and Kupferman, O. 2002. Alternating-time temporal logic. *J. ACM* 49(5):672–713.

Aminof, B.; Kwiatkowska, M.; Maubert, B.; Murano, A.; and Rubin, S. 2019. Probabilistic strategy logic. In Kraus, S., ed., *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019*, 32–38.

Belnap, N., and Perloff, M. 1993. In the realm of agents. *Annals of Mathematics and Artificial Intelligence* 9(1):25–48.

Bulling, N., and Dastani, M. 2013. Coalitional responsibility in strategic settings. In Leite, J.; Son, T. C.; Torroni, P.; van der Torre, L.; and Woltran, S., eds., *Computational Logic in Multi-Agent Systems*, 172–189. Springer Berlin Heidelberg.

Bulling, N., and Jamroga, W. 2009. What agents can probably enforce. *Fundam. Informaticae* 93(1-3):81–96.

Chockler, H., and Halpern, J. Y. 2004. Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research* 20:93–115.

Cunningham, S. 2008. *Driving offences: law, policy and practice*. Routledge.

Dastani, M., and Yazdanpanah, V. 2023. Responsibility of AI systems. *AI & SOCIETY* 38(2):843–852.

Dautović, Š.; Doder, D.; and Ognjanović, Z. 2021. An epistemic probabilistic logic with conditional probabilities. In Faber, W.; Friedrich, G.; Gebser, M.; and Morak, M., eds., *Logics in Artificial Intelligence*, 279–293. Cham: Springer International Publishing.

De Giacomo, G., and Lenzerini, M. 1994. Boosting the correspondence between description logics and propositional dynamic logics. In *Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence*, AAAI'94, 205–212. AAAI Press.

de Lima, T.; Royakkers, L.; and Dignum, F. 2010. A logic for reasoning about responsibility. *Logic Journal of the IGPL* 18(1):99–117.

Deuser, K., and Naumov, P. 2021. Strategic knowledge acquisition. *ACM Transactions on Computational Logic (TOCL)* 22(3):1–18.

Dignum, V. 2019. *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Springer Nature.

Fagin, R., and Halpern, J. Y. 1994. Reasoning about knowledge and probability. *J. ACM* 41(2):340–367.

Feinberg, J. 1984. *The Moral Limits of the Criminal Law. Harm to Others. Volume 1*. Oxford Univ. Press.

Fervari, R.; Herzig, A.; Li, Y.; and Wang, Y. 2017. Strategically knowing how. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 1031–1038.

Frankfurt, H. G. 1969. Alternate possibilities and moral responsibility. *The Journal of Philosophy* 66(23):829–839.

Goranko, V., and Jamroga, W. J. 2004. Comparing semantics of logics for multi-agent systems. *Synthese* 139(2):241–280. Imported from HMI.

Goranko, V.; Jamroga, W.; and Turrini, P. 2013. Strategic games and truly playable effectivity functions. *Agent Multi-Agent Syst* 26:288—314.

Halpern, J. Y., and Kleiman-Weiner, M. 2018. Towards formal definitions of blameworthiness, intention, and moral responsibility. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, AAAI-18, 1853–1860.

Heifetz, A., and Mongin, P. 2001. Probability logic for type spaces. *Games and economic behavior* 35(1-2):31–53.

Herzig, A., and Troquard, N. 2006. Knowing how to play: Uniform choices in logics of agency. In *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems*, AAMAS '06, 209–216. New York, NY, USA: Association for Computing Machinery.

Huang, X.; Su, K.; and Zhang, C. 2012. Probabilistic alternating-time temporal logic of incomplete information and synchronous perfect recall. In Hoffmann, J., and Selman, B., eds., *Proceedings of the Twenty-Sixth AAAI*. AAAI Press.

Jamroga, W., and van der Hoek, W. 2004. Agents that know how to play. *Fundamenta Informaticae* 63(2-3):185–219.

Jamroga, W. 2003. Some remarks on alternating temporal epistemic logic. In Dunin-Keplicz, B., and Verbrugge, R., eds., *Proceedings of Formal Approaches to Multi-Agent Systems (FAMAS 2003)*, 133–139.

Lindahl, L. 1977. *Position and Change: A Study in Law and Logic*. Dordrecht: Springer Netherlands.

Lorini, E.; Longin, D.; and Mayor, E. 2013. A logical analysis of responsibility attribution: emotions, individuals and collectives. *Journal of Logic and Computation* 24(6):1313–1339.

Naumov, P., and Tao, J. 2019. Knowing-how under uncertainty. *Artificial Intelligence* 276:41–56.

Naumov, P., and Tao, J. 2020. An epistemic logic of blameworthiness. *Artificial Intelligence* 283:103269.

Novák, P., and Jamroga, W. 2011. Agents, actions and goals in dynamic environments. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

Parikh, R. 1978. The completeness of propositional dynamic logic. In Winkowski, J., ed., *Mathematical Foundations of Computer Science 1978*, 403–415. Springer Berlin Heidelberg.

Pauly, M. 2002. A Modal Logic for Coalitional Power in Games. *Journal of Logic and Computation* 12(1):149–166.

Ramírez Abarca, A. I., and Broersen, J. 2021. Stit semantics for epistemic notions based on information disclosure in interactive settings. *Journal of Logical and Algebraic Methods in Programming* 123:100708.

Royakkers, L., and Hughes, J. 2020. Blame it on me. *Journal of Philosophical Logic* 49(2):315–349.

Schild, K. 1991. A correspondence theory for terminological logics: Preliminary report. In *International Joint Conference on Artificial Intelligence*.

Sim, J. 2019. *Parents Killing Children: Crossing the Invisible Line*. Routledge.

Smith, H. 2020. Clinical ai: opacity, accountability, responsibility and liability. *AI & SOCIETY* 1–11.

Sommerville, I.; Cliff, D.; Calinescu, R.; Keen, J.; Kelly, T.; Kwiatkowska, M.; Mcdermid, J.; and Paige, R. 2012. Large-scale complex it systems. *Communication of the ACM* 55(7):71–77.

ten Voorde, J. 2014. Prohibiting remote harms: On endangerment, citizenship and control. *Utrecht Law Review* 10:163–179.

van de Poel, I. 2011. The relation between forward-looking and backward-looking responsibility. In Vincent, N. A.; van de Poel, I.; and van den Hoven, J., eds., *Moral Responsibility: Beyond Free Will and Determinism*, 37–52. Dordrecht: Springer Netherlands.

Yazdanpanah, V., and Dastani, M. 2016. Distant group responsibility in multi-agent systems. In Baldoni, M.; Chopra, A. K.; Son, T. C.; Hirayama, K.; and Torroni, P., eds., *PRIMA 2016: Principles and Practice of Multi-Agent Systems*, 261–278. Cham: Springer International Publishing.

Yazdanpanah, V.; Dastani, M.; Jamroga, W.; Alechina, N.; and Logan, B. 2019. Strategic responsibility under imperfect information. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, 592–600. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.