The Limits of Morality in Strategic Games

Rui Cao[1*] and **Pavel Naumov**[2]
[1]University of British Columbia, Canada
[2]University of Southampton, the United Kingdom
rui.cao@sauder.ubc.ca, p.naumov@soton.ac.uk

Abstract

An agent, or a coalition of agents, is blameable for an outcome if she had a strategy to prevent it. In this paper we introduce a notion of limited blameworthiness, with a constraint on the amount of sacrifice required to prevent the outcome. The main technical contribution is a sound and complete logical system for reasoning about limited blameworthiness in the strategic game setting.

Introduction

With humans delegating more and more decision power to autonomous systems such as self-driving cars, automated stock traders, and war robots, there is a need to adapt the notion of responsibility which would be applicable to artificial agents. In addition, autonomous agents must be able to reason about their own and human responsibility in a hybrid human-machine environment. Towards this goal, in this paper we study the responsibility of agents and coalitions of agents in a strategic game setting.

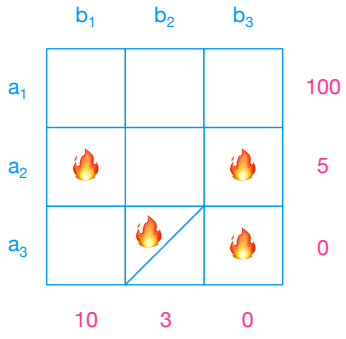

Figure 1: Strategic Game.

As an example, consider the strategic game depicted in Figure 1. This game has two players that we refer to as Alice and Bob. In this game, Alice has actions a_1, a_2, and a_3, while Bob has actions b_1, b_2, and b_3. The outcome of each action profile is depicted in the corresponding cell of the table. For example, if Alice chooses action a_2 while Bob picks b_3, fire starts. Note that we allow games to be nondeterministic. For example, if Alice and Bob choose strategies a_3 and b_2, respectively, then the fire might start or it might not. In this case we say that action profile (a_3, b_2) has two different outcomes: "fire" and "no fire" depicted as half-cells of the tables.

We say that an agent (or a coalition) is *blameable* for φ if statement φ is true and the agent (or the coalition) had a strategy to prevent φ. This definition is often referred to as Frankfurt's [1969] principle of alternative possibilities [Widerker, 2017]. In our example, if Alice and Bob choose actions a_2 and b_3 respectively, then fire starts. In this case, Alice is blamable for the fire:

$$(a_2, b_3, \text{fire}) \Vdash \mathsf{B}_{\text{Alice}}\text{"Fire started"}$$

because fire happened and Alice had a strategy (use action a_1) that would prevent the statement "Fire started" from being true. In the same situation, Bob is not blamable for the start of the fire because he had no strategy to prevent it:

$$(a_2, b_3, \text{fire}) \Vdash \neg\mathsf{B}_{\text{Bob}}\text{"Fire started"}.$$

Note that according to the principle of alternative possibilities, in the case of the action profile (a_3, b_2), whether Alice is blamable or not depends on the outcome of the actions:

$$(a_3, b_2, \text{fire}) \Vdash \mathsf{B}_{\text{Alice}}\text{"Fire started"},$$
$$(a_3, b_2, \text{no fire}) \Vdash \neg\mathsf{B}_{\text{Alice}}\text{"Fire started"}.$$

Naumov and Tao proposed complete logical systems that describe the properties of blameworthiness as a modality in perfect information strategic games [2019], imperfect information strategic games [2020c; 2020b], and security games (2020a).

In many real world situations humans are not blamed for the outcome if preventing it would require sacrificing too much. Can you shoot your neighbour's dog if it is attacking your child? Should you share your house with homeless people? Should you donate $1000 for a good cause? Can you use a cell phone in a movie theater if you are having a heart attack? Should you lend your calculator during an exam if you need it too? Should you proof-read a research paper for your friend? Do you have to finish a job assignment if it requires you to work overtime? Should you climb on a roof to save a kitten?

In the words of Shelly Kagan in *The Limits of Morality*,

> The greater the sacrifice which morality requires, obviously enough, the more significantly it will decrease the agent's ability to mold his life as he

*Contact Author

> chooses and to promote his interests. The moderate may want to argue that more than a certain loss of such autonomy is morally intolerable. Hence, morality can exact so much, and no more. [Kagan, 1991, p.21]

In this paper we consider limited blameworthiness modality $\mathsf{B}^s_a\varphi$ which means that formula φ is true and agent a had a strategy with cost at most s to prevent it. Let us return to our example depicted in Figure 1 and to assume that the cost of each action for each agent is as shown in the figure. For example, the cost of action a_1, that Alice can use to prevent the fire, is 100. Thus, if Alice chooses action a_2, Bob chooses b_3, and fire starts, then Alice can be blamed for the fire if the moral limit on her sacrifice is 100

$$(a_2, b_3, \text{fire}) \Vdash \mathsf{B}^{100}_{\text{Alice}} \text{“Fire started”}$$

because she has an action with cost at most 100 to prevent the fire. At the same time, if the moral limit on Alice's sacrifice is set to 50, then she cannot be blamed for the fire

$$(a_2, b_3, \text{fire}) \Vdash \neg\mathsf{B}^{50}_{\text{Alice}} \text{“Fire started”}$$

because she has no action with cost at most 50 to prevent the fire. Finally, note that, as a coalition, Alice and Bob can prevent the fire if Alice chooses action a_2 and Bob chooses action b_2 with total cost $5 + 3 = 8$. Thus, we say that their coalition is blameable with sacrifice 8,

$$(a_2, b_3, \text{fire}) \Vdash \mathsf{B}^{8}_{\text{Alice, Bob}} \text{“Fire started”}.$$

In this paper we propose a sound and complete logical system that captures the universal properties of limited blameworthiness modality $\mathsf{B}^s_C\varphi$. The axioms of the proposed systems are variations of the axioms for blameworthiness modality $\mathsf{B}_C\varphi$ without sacrifice limit [Naumov and Tao, 2019]. However, the proof of the completeness in the current paper is very different from [Naumov and Tao, 2019]; it is closer to the proof of the completeness for Resource-Bounded Coalition Logic [Alechina *et al.*, 2011]. We further discuss the relation between these works in the beginning of Completeness section.

Other Related Literature

The other form of responsibility discussed in the literature is responsibility for *seeing to* φ. An agent is responsible for seeing to φ if the action taken by the agent unavoidably leads to φ being true. This notion of responsibility is captured in the logic of seeing-to-it-that (STIT) [Belnap and Perloff, 1990; Horty, 2001; Horty and Belnap, 1995; Horty and Pacuit, 2017; Olkhovikov and Wansing, 2018]. The cost of sacrifice could be potentially added to responsibility for seeing to it and interpreted as a *degree of praiseworthiness*. In many situation, the more an agent had to sacrifice to see to φ, the more she should be praised for φ.

Naumov and Yew proposed a *dilemma modality* that captures a hard choice that an agent faces between two or more undesirable alternatives [2021]. Although this modality in itself does not represent a form of responsibility, it also contains a cost of sacrifice. In their case, the sacrifice is the limit on the cost of actions that eliminates costly actions that could have been used to avoid making the hard choice.

Strategic Games with Cost of Actions

In this section, we describe the class of games that will be used as a semantics of our logical system. Throughout this paper, we assume a fixed finite set $\mathcal{A}$ of agents and a fixed set of propositional variables. By a coalition we mean an arbitrary subset of $\mathcal{A}$. By X^Y we denote the set of all functions from set Y to set X.

Definition 1. *A game is a tuple* $(\Delta, \|\cdot\|, d_0, \Omega, P, \pi)$*, where*

1. Δ *is a nonempty set of “actions”, any function from set* $\Delta^{\mathcal{A}}$ *is called a “complete action profile”,*
2. $\|d\|$ *is a nonnegative real number, called the cost of action* $d \in \Delta$*,*
3. $d_0 \in \Delta$ *is a zero-cost action:* $\|d_0\| = 0$*,*
4. Ω *is a set of “outcomes”,*
5. *a set of “plays”* P *is an arbitrary subset of* $\Delta^{\mathcal{A}} \times \Omega$ *satisfying the following* ***nontermination*** *condition: for each complete action profile* $\delta \in \Delta^{\mathcal{A}}$ *there is at least one outcome* $\omega \in \Omega$ *such that* $(\delta, \omega) \in P$*,*
6. π *is a function that maps propositional variables into subsets of* P*.*

In the introductory example, Alice and Bob had different sets of actions. In Definition 1, we assume that each agent has the same set of actions Δ. This assumption is not significant. We made it to simplify the notations. We interpret the cost of any action $\|d\|$ as the amount of “sacrifice” that this action requires. We will reflect on the reasons for considering only nonnegative sacrifice in Syntax and Semantics section when we discuss sacrifice of a coalition. Although in this paper we assume that the cost of an action does not depend on the agent, this restriction is also not significant.

In game theory, strategic games are usually assumed to be deterministic in the sense that the outcomes (pay-offs) are uniquely determined by the complete action profile. Nondeterminacy in such games is usually modeled through an additional player, often interpreted as “nature” or “god”. In this paper we have chosen to model nondeterminacy by explicitly assuming that a complete action profile might correspond to multiple outcomes. In addition to simplifying our axiomatic system, this choice also avoids the necessity to assign blame to the “nature”.

The game depicted in Figure 1 has 9 complete action profiles corresponding to possible combinations of Alice's and Bob's strategies. This game has 10 outcomes defined by the “regions” of the table (eight squares and 2 triangles). In that game, each outcome corresponds to a unique complete action profile. In general, it will be convenient to assume that different complete action profiles might result in the same outcome. Thus, we introduce the set of “plays” P that specifies which complete action profiles are consistent with which outcomes. Informally, this set captures the rules of the game. The nontermination condition requires each complete action profile to lead to at least one outcome.

The existence of a zero-cost action in the game is important for capturing the notion of limited blameworthiness. Indeed, consider an agent that has a choice of just two actions: not to save somebody's life at cost 99 or to save the life at cost 100.

Intuitively, the sacrifice in such a setting is 1, not 100. By requiring that there exists a zero-cost action, we, in essence, "normalize" the costs of all actions available to the agent.

Recall from the introductory examples that we place a play, not a state, on the left side of $\Vdash$. In other words, statements in our logical systems are not about states, but about plays. Similarly, we interpret propositional variables as statements about plays. This is why valuation function π in Definition 1 maps propositional variables into sets of plays rather than sets of outcomes.

Definition 2. *For any action profile $\gamma \in \Delta^C$ of a coalition C, by $\|\gamma\|$ we mean the total cost of the action profile to the coalition: $\|\gamma\| = \sum_{a \in C} \|\gamma(a)\|$.*

By defining the cost to a coalition as a sum of the costs of actions of the individual members of the coalition, we ignore the distribution of the burden between the individual members. Although this makes sense in many real-world situations, it is also easy to see that in some cases this approach might be problematic. For example, one probably should blame a coalition of 10 people for breaking a rule if it cost only \$19 to all of them together to follow the rule. Indeed, it is less than \$2 per person. Similarly, one might argue that 1,000,000 people should be blamed for breaking the rule if it costs them total of \$1,999,999 to follow the rule. But what if it costs \$1,000,000 to one person and just \$1 to each of the remaining 999,999 people?

In this paper we decided to disallow negative costs of actions because they lead to even more paradoxical situations. For example, consider a situation when to prevent φ one needs a joint effort of Alice and Bob. Suppose that the action required of Alice will cost her \$1,000,000, while Bob will make a profit (negative cost) on his action in the amount of \$1,000,000. Should we blame them, as a coalition, for not preventing φ at *zero cost*?

Syntax and Semantics

The language Φ of our system is defined by the grammar

$$\varphi := p \mid \neg\varphi \mid \varphi \to \varphi \mid \mathsf{N}\varphi \mid \mathsf{B}^s_C\varphi,$$

where p is a propositional variable, $s \geq 0$ is a real number, and C is a coalition. We read $\mathsf{B}^s_C\varphi$ as "coalition C is blameable for statement φ because it would have to sacrifice at most s to prevent φ". We read $\mathsf{N}\varphi$ as "statement φ is universally true in the given game". By $\overline{\mathsf{N}}\varphi$ we denote formula $\neg\mathsf{N}\neg\varphi$. Thus, $\overline{\mathsf{N}}\varphi$ means that statement φ is true on at least one play of the game. We assume that conjunction $\wedge$, disjunction $\vee$, biconditional $\leftrightarrow$, and truth constant $\top$ are defined as usual.

For any functions f and g, we write $f =_X g$, if $f(x) = g(x)$ for each $x \in X$. The following is the key definition of this paper. Item 5 of this definition formally specifies the meaning of the blameworthiness modality $\mathsf{B}^s_C\varphi$.

Definition 3. *For any formula φ in language Φ and any play $(\delta, \omega) \in P$ of a game $(\Delta, \|\cdot\|, d_0, \Omega, P, \pi)$, the satisfaction relation $(\delta, \omega) \Vdash \varphi$ is defined recursively as follows:*

1. *$(\delta, \omega) \Vdash p$ if $(\delta, \omega) \in \pi(p)$, where p is a propositional variable,*
2. *$(\delta, \omega) \Vdash \neg\varphi$ if $(\delta, \omega) \nVdash \varphi$,*
3. *$(\delta, \omega) \Vdash \varphi \to \psi$ if $(\delta, \omega) \nVdash \varphi$ or $(\delta, \omega) \Vdash \psi$,*
4. *$(\delta, \omega) \Vdash \mathsf{N}\varphi$ if $(\delta', \omega') \Vdash \varphi$ for each play $(\delta', \omega') \in P$,*
5. *$(\delta, \omega) \Vdash \mathsf{B}^s_C\varphi$ if*
   (a) *$(\delta, \omega) \Vdash \varphi$,*
   (b) *there is a profile $\gamma \in \Delta^C$ such that $\|\gamma\| \leq s$ and for each play $(\delta', \omega') \in P$, if $\gamma =_C \delta'$, then $(\delta', \omega') \nVdash \varphi$,*
   (c) *for each proper subset D of set C and each action profile $\gamma \in \Delta^D$ where $\|\gamma\| \leq s$, there is a play $(\delta', \omega') \in P$ such that $\gamma =_D \delta'$ and $(\delta', \omega') \Vdash \varphi$.*

Note that if a coalition had a strategy to prevent condition φ, then every superset of the coalition had such strategy as well. To avoid blaming a coalition for a failure of its subset, the above definition includes the minimality condition captured by item 5(c). Namely, we say that a coalition C is blamable for statement φ with sacrifice limit s if φ is true, coalition C had a strategy to prevent it at cost no more than s, and no subset of C had a strategy to prevent φ at cost no more than s.

Although the intended meaning of the sacrifice s in modal formula $\mathsf{B}^s_C\varphi$ is to capture the *limit* of the blameworthiness, it can also be interpreted as a *degree* of blameworthiness. Intuitively, the lower the cost of prevention is, the more blamable the coalition should be for the outcome. Halpern and Kleiman-Weiner use the cost of prevention to define one of the degrees of blameworthiness that they propose (2018). An important difference between Definition 3 and Halpern and Kleiman-Weiner's approach is that we use absolute sacrifice while they use relative sacrifice. Relative sacrifice is the difference between the current costs encountered by a coalition and the costs required to prevent the undesirable outcome. For example, consider a business company that spent \$100,000 on a safety device that does not completely prevent a certain type of accident, but the company could have spent \$100,001 on a more expensive version of the device that completely prevents this type of accident. If the accident happens, should the degree of blame be computed based on absolute cost \$100,001 or the relative cost of \$1? In the former case the degree of blame will be low, in the second case it will be high. In this example, it probably makes sense to use the relative sacrifice as proposed in [Halpern and Kleiman-Weiner, 2018]. However, if in the same setting the original \$100,000 were spent not on a safety device, but on a new warehouse, then the absolute sacrifice of \$100,001 is probably a better measure of the degree of company's blameworthiness for the accident. The logical system proposed in this paper is suitable for reasoning about the latter, but not the former setting.

Finally, the discussed above degree of blameworthiness *of a group* should not be confused with the degree of responsibility *within the group*. The latter is concerned with how the blame should be divided between several members of the group. In the spirit of this paper, one might suggest that degree within the group could be defined as a ratio of total costs that would be imposed on the whole group to the cost imposed on an individual. For example, if prevention would require a joint effort of Alice and Bob and the cost of the appropriate actions is \$2 for Alice and \$5 for Bob, then Alice

is responsible for the outcome to a higher degree than Bob. However, this approach is problematic because there might be two different ways to prevent outcome: one of them would cost \$2 to Alice and \$5 to Bob, while the other would cost \$5 to Alice and \$2 to Bob. Thus, the cost-based approach explored in this paper in some situations might be used to define the degree of blameworthiness of a group but it is not likely to be appropriate to define the degree of responsibility within the group.

In this paper we use abbreviation $\mathbb{B}^s_C\varphi$ for the disjunction $\bigvee_{D\subseteq C} \mathsf{B}^s_D\varphi$. Informally, $\mathbb{B}^s_C\varphi$ means "statement φ is true and one of subsets of C could be blamed for it with sacrifice s". As shown in the full version of the paper [Cao and Naumov, 2019], it is equivalent to statement "φ is true and coalition C could have prevented it at cost no more than s".

Axioms

In addition to the propositional tautologies in language Φ, our logical system contains the following axioms:

1. Truth: $\mathsf{N}\varphi \to \varphi$ and $\mathsf{B}^s_C\varphi \to \varphi$,
2. Distributivity: $\mathsf{N}(\varphi \to \psi) \to (\mathsf{N}\varphi \to \mathsf{N}\psi)$,
3. Euclidicity: $\neg\mathsf{N}\varphi \to \mathsf{N}\neg\mathsf{N}\varphi$,
4. None to Blame: $\neg\mathsf{B}^s_\varnothing\varphi$,
5. Blamelessness of Truth: $\neg\mathsf{B}^s_C\top$,
6. Monotonicity: $\mathsf{B}^s_C\varphi \to \mathbb{B}^t_C\varphi$, where $s \le t$,
7. Minimality: $\mathsf{B}^s_C\varphi \to \neg\mathsf{B}^s_D\varphi$, where $D \subsetneq C$,
8. Joint Responsibility: if $C \cap D = \varnothing$, then $\overline{\mathsf{N}}\mathsf{B}^s_C\varphi \wedge \overline{\mathsf{N}}\mathsf{B}^t_D\psi \to (\varphi \vee \psi \to \mathbb{B}^{s+t}_{C\cup D}(\varphi \vee \psi))$,
9. Strict Conditional: $\mathsf{N}(\varphi \to \psi) \to (\mathsf{B}^s_C\psi \to (\varphi \to \mathbb{B}^s_C\varphi))$,
10. Fairness: $\mathsf{B}^s_C\varphi \to \mathsf{N}(\varphi \to \mathsf{B}^s_C\varphi)$,
11. Substitution: $\mathsf{N}(\varphi \leftrightarrow \psi) \to (\mathsf{B}^s_C\varphi \to \mathsf{B}^s_C\psi)$.

The Truth, the Distributivity, and the Euclidicity axioms for modality N capture the fact that this is an S5-modality [Fagin *et al.*, 1995]. The Truth axiom for modality B states that a coalition can only be blamed for something which is true. The None to Blame axiom says that the empty coalition can not be blamed for anything. The Blamelessness of Truth axioms says that none can be blamed for a tautology. Informally, this axiom is sound because there could be no strategy to prevent a tautology. The Monotonicity axiom states that if a coalition C can be blamed for not preventing an outcome at cost at most s, then either the coalition itself or at least one of its subsets can also be blamed for not preventing the outcome at cost at most t, where $t \ge s$. The Minimality axiom says that if a coalition can be blamed for φ, then no proper subset of this coalition can be blamed for φ. The Joint Responsibility axiom shows how blames of two disjoint coalitions can be combined into a blame of their union. To understand the Strict Conditional axiom, note that formula $\mathsf{N}(\varphi \to \psi)$ means that φ implies ψ for each play of the game. The axiom says that if a coalition is responsible for statement ψ, then either the coalition itself or at least one of its subsets is responsible for a stronger statement φ as long as φ is true. The Fairness axiom states that if a coalition is blamed for φ, then it should be blamed for φ each time when φ is true.

We write $\vdash \varphi$ if formula φ is provable from the axioms of our system using the Modus Ponens and the Necessitation inference rules:

$$\frac{\varphi, \varphi \to \psi}{\psi}, \qquad \frac{\varphi}{\mathsf{N}\varphi}.$$

If $\vdash \varphi$, then we say that formula φ is a theorem of our logical system. In addition to unary relation $\vdash \varphi$, we also consider a binary relation $X \vdash \varphi$. We write $X \vdash \varphi$ if formula φ is provable from all *theorems* of our logical system and the set of additional formulae X using the Modus Ponens inference rule only. It is easy to see that statement $\varnothing \vdash \varphi$ is equivalent to $\vdash \varphi$. A set of formulae X is consistent if there is no formula φ such that $X \vdash \varphi$ and $X \vdash \neg\varphi$.

Lemma 1 (Lindenbaum)**.** *Any consistent set of formulae can be extended to a maximal consistent set of formulae.*

Proof. The standard proof of Lindenbaum's lemma applies here [Mendelson, 2009, Proposition 2.14]. However, since the formulae in our logical system use real numbers in superscript, the set of formulae is uncountable. Thus, the proof of Lindenbaum's lemma in our case relies on the Axiom of Choice. ⊠

Completeness

We show the soundness of our logical system in the full version of this paper [Cao and Naumov, 2019]. In the rest of this paper, we prove its completeness.

We will use modality $\Box^s_C\varphi$ as an abbreviation for statement $\mathsf{N}(\neg\varphi \to \mathbb{B}^s_C\neg\varphi)$. Informally, formula $\Box^s_C\varphi$ means that coalition C or one of its subsets is blamable for $\neg\varphi$ in each outcome of the game in which statement φ is false. In other words, $\Box^s_C\varphi$ is a counterfactual modality that means that coalition C *could have* prevented $\neg\varphi$ at cost s. This modality is different from the "coalition C has a strategy to achieve φ at costs s" modality of Resource-Bounded Coalition Logic (RBCL) [Alechina *et al.*, 2011] denoted here by $\mathsf{S}^s_C\varphi$. For example, statement $\varphi \wedge \Box^s_C\neg\varphi$ means that φ is true now, but coalition C could have prevented it at cost s. At the same time, statement $\varphi \wedge \mathsf{S}^s_C\neg\varphi$ in RBCL means that φ is true now but C could make it false in the future. In spite of this semantical difference between modalities $\Box^s_C\varphi$ and $\mathsf{S}^s_C\varphi$, they share many common properties. In the full version of this paper [Cao and Naumov, 2019], we prove the following basic properties of the modalities $\Box^s_C\varphi$ and N. Some of these properties are listed as axioms in [Alechina *et al.*, 2011].

P0. $\vdash \Box^s_C\varphi \to \mathsf{N}\Box^s_C\varphi$.

P1. If $\varphi_1, \dots, \varphi_n \vdash \psi$, $t_1 + \dots + t_n \le s$, and $D_1, \dots, D_n$ are pairwise disjoint subsets of set C, then $\Box^{t_1}_{D_1}\varphi_1, \dots, \Box^{t_n}_{D_n}\varphi_n \vdash \Box^s_C\psi$,

P2. $\vdash \mathsf{N}\varphi \to \Box^s_C\varphi$,

P3. $\vdash \Box^t_\varnothing\varphi \to \mathsf{N}\varphi$,

P4. $\vdash \neg\Box^s_C\neg\top$.

P5. $\vdash \mathsf{N}\varphi \to \mathsf{N}\mathsf{N}\varphi$,

P6. $\varphi_1, \ldots, \varphi_n \vdash \psi$, then $\mathsf{N}\varphi_1, \ldots, \mathsf{N}\varphi_n \vdash \mathsf{N}\psi$,

P7. $\vdash \neg \mathsf{B}^s_C \varphi \to \neg\varphi \vee \neg\Box^s_C \neg\varphi \vee \bigvee_{D \subsetneq C} \Box^s_D \neg\varphi$.

P8. $\vdash \mathsf{B}^s_C \varphi \to \Box^s_C \neg\varphi$.

P9. $\vdash \mathsf{B}^s_C \varphi \to \neg\Box^s_D \neg\varphi$, where $D \subsetneq C$.

In the rest of the paper we use these properties to prove the completeness of our logical system.

Compared to this paper, the blameworthiness modality in [Naumov and Tao, 2019] is missing not only the sacrifice constraint, but also the minimality condition of item 5(c) in Definition 3. As a result, the blameworthiness modality in [Naumov and Tao, 2019] is, essentially, a sacrifice-free version of modality $\mathbb{B}$. The proof of the completeness in [Naumov and Tao, 2019] does not introduce anything similar to abbreviations $\Box^s_C \varphi$. Instead, it constructs the canonical game using the blameworthiness modality directly.

Canonical Model

As usual in modal logic, the proof of the completeness relies on the construction of a canonical model. In our case, we define the canonical game $G(\omega_0) = (\Delta, \|\cdot\|, d_0, \Omega, P, \pi)$ for each maximal consistent set of formulae ω_0. We define each component of the canonical game $G(\omega_0)$ separately.

Definition 4. *Set Δ consists of a zero-cost action d_0, which is not a triple, and all triples (φ, C, s) such that $\varphi \in \Phi$ is a formula, C is a nonempty coalition, and s is a non-negative real number.*

Informally, we consider actions as "votes" of agents. Zero-cost action d_0 could be interpreted as abstaining from voting. Action (φ, C, s) by an agent a means that agent a is voting as a part of coalition C to force φ at the total cost s to the whole coalition. If agent a votes (φ, C, s), then statement φ is not necessarily true in the outcome. The vote aggregation mechanism is given in Definition 7. Definition 4 is substantially different from a similar definition in [Naumov and Tao, 2019], where each action consists of just a single formula φ.

Definition 5. *For each action $d \in \Delta$, let $\|d\| = 0$ if $d = d_0$ and $\|d\| = \dfrac{s}{|C|}$ if $d = (\varphi, C, s)$.*

Informally, $\|d\| = \frac{s}{|C|}$ means that the cost of each joint action is divided evenly between all members of the coalition. Note that size $|C|$ of coalition C is non-zero by Definition 4.

Definition 6. *The set of outcomes Ω is the set of all maximal consistent sets of formulae ω such that for each formula φ if $\mathsf{N}\varphi \in \omega_0$, then $\varphi \in \omega$.*

Definition 7. *The set $P \subseteq \Delta^{\mathcal{A}} \times \Omega$ consists of all pairs (δ, ω) such that for any $\Box^s_C \psi \in \omega_0$, if $\delta(a) = (\psi, C, s)$ for each agent $a \in C$, then $\psi \in \omega$.*

In other words, for each formula $\Box^s_C \psi \in \omega_0$, if each member of coalition C votes as a part of C to force ψ at cost s, then ψ is guaranteed to be true in the outcome.

Definition 8. *$\pi(p) = \{(\delta, \omega) \in P \mid p \in \omega\}$ for each propositional variable p.*

This concludes the definition of the canonical game $G(\omega_0)$. In Lemma 5, we prove the nontermination condition from item 5 of Definition 1 is satisfied for game $G(\omega_0)$.

As usual, the key part of the proof of the completeness is the induction, or "truth", lemma. In our case this is Lemma 7. The next three lemmas are auxiliary lemmas used in the proof of Lemma 7.

Lemma 2. *For any play $(\delta, \omega) \in P$ and any formula $\Box^s_C \neg\varphi \in \omega$ there is a profile $\gamma \in \Delta^C$ where $\|\gamma\| \leq s$ such that for any play $(\delta', \omega') \in P$ if $\gamma =_C \delta'$, then $\varphi \notin \omega'$.*

Proof. Define $\gamma \in \Delta^C$ to be an action profile of coalition C such that for each agent $a \in C$,

$$\gamma(a) = (\neg\varphi, C, s). \tag{1}$$

Claim 1. $\|\gamma\| \leq s$.

PROOF OF CLAIM. If set C is not empty, then, by Definition 2 and Definition 5, $\|\gamma\| = \sum_{a \in C} \|(\varphi, C, s)\| = \sum_{a \in C} \frac{s}{|C|} = s$.

If set C is empty, then $\|\gamma\| = 0$ by Definition 2. At the same time, $s \geq 0$ by the definition of language Φ. Therefore, $\|\gamma\| \leq s$. ⊠

Consider any play $(\delta', \omega') \in P$ such that $\gamma =_C \delta'$. Recall that $\Box^s_C \neg\varphi \in \omega$ by the assumption of the lemma. Hence $\omega \vdash \mathsf{N}\Box^s_C \neg\varphi$ by Property P0. Thus, $\neg\mathsf{N}\Box^s_C \neg\varphi \notin \omega$ because set ω is maximal. Then, $\mathsf{N}\neg\mathsf{N}\Box^s_C \neg\varphi \notin \omega_0$ by Definition 6. Hence, because set ω_0 is maximal,

$$\neg\mathsf{N}\neg\mathsf{N}\Box^s_C \neg\varphi \in \omega_0. \tag{2}$$

At the same time, formula $\neg\mathsf{N}\Box^s_C \neg\varphi \to \mathsf{N}\neg\mathsf{N}\Box^s_C \neg\varphi$ is an instance of the Euclidicity axiom. Thus, by contraposition, $\vdash \neg\mathsf{N}\neg\mathsf{N}\Box^s_C \neg\varphi \to \mathsf{N}\Box^s_C \neg\varphi$. Hence, by the Truth axiom and the propositional reasoning, $\vdash \neg\mathsf{N}\neg\mathsf{N}\Box^s_C \neg\varphi \to \Box^s_C \neg\varphi$. Then, $\omega_0 \vdash \Box^s_C \neg\varphi$ by the Modus Ponens inference rule using statement (2). Hence, $\Box^s_C \neg\varphi \in \omega_0$ because set ω_0 is maximal. Thus, $\neg\varphi \in \omega'$ by Definition 7, the assumption $\gamma =_C \delta'$ and statement (1). Then, $\varphi \notin \omega'$ because ω' is consistent. ⊠

Lemma 3. *For any play $(\delta, \omega) \in P$, any profile $\gamma \in \Delta^C$, and any formula $\neg\Box^s_C \neg\varphi \in \omega$, if $\|\gamma\| \leq s$, then there is a play $(\delta', \omega') \in P$ such that $\gamma =_C \delta'$ and $\varphi \in \omega'$.*

Proof. Consider the following set X of formulae:

$$\{\varphi\} \cup \{\psi \mid \mathsf{N}\psi \in \omega_0\} \cup \{\chi \mid \Box^t_D \chi \in \omega_0, D \subseteq C, \forall a \in D(\gamma(a) = (\chi, D, t))\}.$$

Claim 2. *Set X is consistent.*

PROOF OF CLAIM. Suppose the opposite. Thus, there are

$$\mathsf{N}\psi_1, \ldots, \mathsf{N}\psi_m, \Box^{t_1}_{D_1} \chi_1, \ldots, \Box^{t_n}_{D_n} \chi_n \in \omega_0, \tag{3}$$

such that

$$D_1, \ldots, D_n \subseteq C, \tag{4}$$

$$\gamma(a) = (\chi_i, D_i, t_i) \text{ for all } a \in D_i, i \leq n, \tag{5}$$

$$\psi_1, \ldots, \psi_m, \chi_1, \ldots, \chi_n \vdash \neg\varphi. \tag{6}$$

Without loss of generality, we can assume that formulae $\chi_1, \ldots, \chi_n$ are distinct. Thus, assumption (5) implies that sets $D_1, \ldots, D_n$ are pairwise disjoint. Hence, by Definition 5 and formula (5),

$$\begin{aligned}\|\gamma\| &= \sum_{a\in C}\|\gamma(a)\| \geq \sum_{a\in D_1}\|\gamma(a)\| + \cdots + \sum_{a\in D_n}\|\gamma(a)\| \\ &= \sum_{a\in D_1}\|(\chi_1, D_1, t_1)\| + \cdots + \sum_{a\in D_n}\|(\chi_n, D_n, t_n)\| \\ &= \sum_{a\in D_1}\frac{t_1}{|D_1|} + \cdots + \sum_{a\in D_n}\frac{t_n}{|D_n|} = t_1 + \cdots + t_n.\end{aligned}$$

Thus, $t_1 + \cdots + t_n \leq s$ by the assumption $\|\gamma\| \leq s$ of the lemma. Then, assumption (6) by Property P1 implies

$$\Box^0_\varnothing\psi_1, \ldots, \Box^0_\varnothing\psi_m, \Box^{t_1}_{D_1}\chi_1, \ldots, \Box^{t_n}_{D_n}\chi_n \vdash \Box^s_C\neg\varphi.$$

Hence, by Property P3 and the Modus Ponens rule applied m times, $\mathsf{N}\psi_1, \ldots, \mathsf{N}\psi_m, \Box^{t_1}_{D_1}\chi_1, \ldots, \Box^{t_n}_{D_n}\chi_n \vdash \Box^s_C\neg\varphi$. Thus, $\omega_0 \vdash \Box^s_C\neg\varphi$ due to statement (3). Hence, $\omega_0 \vdash \mathsf{N}\Box^s_C\neg\varphi$ by Property P0 and the Modus Ponens inference rule. Then, $\mathsf{N}\Box^s_C\neg\varphi \in \omega_0$ because set ω_0 is maximal. Thus, $\Box^s_C\neg\varphi \in \omega$ by Definition 6 and because $\omega \in \Omega$. Thus, $\neg\Box^s_C\neg\varphi \notin \omega$ because set ω is consistent, which contradicts assumption $\neg\Box^s_C\neg\varphi \in \omega$ of the lemma. Then, set X is consistent. ⊠

By Lemma 1, set X can be extended to a maximal consistent set ω'. Thus, $\varphi \in X \subseteq \omega'$ by the choice of sets X and ω'. Also, $\omega' \in \Omega$ by Definition 6 and the choice of sets X and ω'.

Let the complete action profile δ' be defined as follows:

$$\delta'(a) = \begin{cases}\gamma(a), & \text{if } a \in C,\\ d_0, & \text{otherwise.}\end{cases} \tag{7}$$

Then, $\gamma =_C \delta'$.

Claim 3. $(\delta', \omega') \in P$.

PROOF OF CLAIM. Consider any formula $\Box^t_D\chi \in \omega_0$ such that $\delta'(a) = (\chi, D, t)$ for each $a \in D$. By Definition 7, it suffices to show that $\chi \in \omega'$.

Case I: $D \subseteq C$. Thus, $\chi \in X$ by the definition of set X. Therefore, $\chi \in \omega'$ by the choice of set ω'.

Case II: $D \nsubseteq C$. Consider any $a \in D \setminus C$. Thus, $\delta'(a) = d_0$ by equation (7). At the same time, $\delta'(a) = (\chi, D, t)$ because $a \in D$. Thus, $d_0 = (\chi, D, t)$, which is a contradiction, because d_0 is not a triple by Definition 4. ⊠

This concludes the proof of the lemma. ⊠

Lemma 4. *For each outcome $\omega \in \Omega$, there is a complete action profile $\delta \in \Delta^{\mathcal{A}}$ such that $(\delta, \omega) \in P$.*

Proof. Consider a complete action profile δ where $\delta(a) = d_0$ for all $a \in \mathcal{A}$. To show $(\delta, \omega) \in P$, consider any such formula $\Box^t_D\chi \in \omega_0$ that $\delta(a) = (\chi, D, t)$ for all $a \in D$. Due to Definition 7, it enough to prove that $\chi \in \omega$.

Case I: $D = \varnothing$. Thus, assumption $\Box^t_D\chi \in \omega_0$ implies $\omega_0 \vdash \mathsf{N}\chi$ by Property P3 and the Modus Ponens inference rule. Hence, $\mathsf{N}\chi \in \omega_0$ because set ω_0 is maximal. Thus, $\chi \in \omega$ by Definition 6 because $\omega \in \Omega$.

Case II: $D \neq \varnothing$. Hence, set D contains at least one agent a. Then, $(\chi, D, t) = \delta(a) = d_0$ by the definition of profile δ. Thus, $d_0 = (\chi, D, t)$, which is a contradiction, because d_0 is not a triple by Definition 4. ⊠

Next, we show that canonical game $G(\omega_0)$ satisfies the nontermination property from item 5 of Definition 1.

Lemma 5. *For any complete action profile $\delta \in \Delta^{\mathcal{A}}$ there is an outcome $\omega \in \Omega$ such that $(\delta, \omega) \in P$.*

Proof. Recall that game $G(\omega_0)$ is defined for a given maximal consistent set ω_0. Then, $\omega_0 \in \Omega$ by Definition 6 and the Truth axiom. Hence, by Lemma 4, there is a complete action profile $\delta_0 \in \Delta^{\mathcal{A}}$ such that $(\delta_0, \omega_0) \in P$.

Note that $\|\delta\|$ is a finite number because the set of all agents $\mathcal{A}$ is finite. Thus, $\vdash \neg\Box^{\|\delta\|}_{\mathcal{A}}\neg\top$ by Property P4. Hence, $\neg\Box^{\|\delta\|}_{\mathcal{A}}\neg\top \in \omega_0$ because set ω_0 is maximal. Thus, by Lemma 3 in the case of the play (δ_0, ω_0), there is a play $(\delta', \omega') \in P$ such that $\delta =_{\mathcal{A}} \delta'$ and $\top \in \omega'$. Note that $\delta =_{\mathcal{A}} \delta'$ implies that complete action profiles δ and δ' are the same. Then, $(\delta, \omega') \in P$. Choose ω to be ω'. ⊠

Lemma 6. *For each play $(\delta, \omega) \in P$ and each formula $\neg\mathsf{N}\varphi \in \omega$, there is a play $(\delta', \omega') \in P$ such that $\neg\varphi \in \omega'$.*

Proof. Let X be the set $\{\neg\varphi\} \cup \{\psi \mid \mathsf{N}\psi \in \omega_0\}$. Next, we prove the consistency of set X. Assume the opposite. Hence, there are formulae $\mathsf{N}\psi_1, \ldots, \mathsf{N}\psi_n \in \omega_0$ where $\psi_1, \ldots, \psi_n \vdash \varphi$. Thus, $\mathsf{N}\psi_1, \ldots, \mathsf{N}\psi_n \vdash \mathsf{N}\varphi$ due to Property P6. Then, $\omega_0 \vdash \mathsf{N}\varphi$ because $\mathsf{N}\psi_1, \ldots, \mathsf{N}\psi_n \in \omega_0$. Thus, $\omega_0 \vdash \mathsf{NN}\varphi$ by Property P5. Hence, it follows from assumption $\omega \in \Omega$ and Definition 6 that $\mathsf{N}\varphi \in \omega$. Thus, $\neg\mathsf{N}\varphi \notin \omega$ by the consistency of set ω, which contradicts the assumption $\neg\mathsf{N}\varphi \in \omega$ of the lemma. Therefore, set X is consistent. By Lemma 1, set X could be extended to a maximal consistent set ω'. Observe that $\neg\varphi \in X \subseteq \omega'$ due to the definition of set X. Finally, by Lemma 4, there is a profile δ' where $(\delta', \omega') \in P$. ⊠

Next is the "truth" lemma for our proof of the completeness theorem, whose proof can be found in the full version of this paper [Cao and Naumov, 2019].

Lemma 7. *$(\delta, \omega) \Vdash \varphi$ iff $\varphi \in \omega$ for any play $(\delta, \omega) \in P$ and any formula $\varphi \in \Phi$.*

Strong Completeness

Theorem 1. *If $X \nvdash \varphi$, then there is a game and a play (δ, ω) of the game where $(\delta, \omega) \Vdash \chi$ for all $\chi \in X$ and $(\delta, \omega) \nVdash \varphi$.*

Proof. Assume that $X \nvdash \varphi$. Hence, set $X \cup \{\neg\varphi\}$ is consistent. By Lemma 1, set $X \cup \{\neg\varphi\}$ can be extended to a maximal consistent set ω_0. Let $G(\omega_0) = (\Delta, \|\cdot\|, d_0, \Omega, P, \pi)$ to be the canonical game defined above. Then, $\omega_0 \in \Omega$ by Definition 6 and the Truth axiom.

By Lemma 4, there exists an action profile $\delta \in \Delta^{\mathcal{A}}$ such that $(\delta, \omega_0) \in P$. Hence, $(\delta, \omega_0) \Vdash \chi$ for all $\chi \in X$ and $(\delta, \omega_0) \Vdash \neg\varphi$ by Lemma 7 and the choice of set ω_0. Therefore, $(\delta, \omega_0) \nVdash \varphi$ by Definition 3. ⊠

Conclusion

In this paper we combine the ideas from the logics of resource bounded coalitions [Alechina *et al.*, 2011] and blameworthiness [Naumov and Tao, 2019] into a logical system that captures the properties of limits of blameworthiness. The main technical result is a strongly complete logical system describing the limited blameworthiness modality.

References

[Alechina *et al.*, 2011] Natasha Alechina, Brian Logan, Hoang Nga Nguyen, and Abdur Rakib. Logic for coalitions with bounded resources. *Journal of Logic and Computation*, 21(6):907–937, December 2011.

[Belnap and Perloff, 1990] Nuel Belnap and Michael Perloff. Seeing to it that: A canonical form for agentives. In *Knowledge representation and defeasible reasoning*, pages 167–190. Springer, 1990.

[Cao and Naumov, 2019] Rui Cao and Pavel Naumov. The limits of morality in strategic games. *arXiv:1901.08467*, 2019.

[Fagin *et al.*, 1995] Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. *Reasoning about knowledge*. MIT Press, Cambridge, MA, 1995.

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

[Halpern and Kleiman-Weiner, 2018] Joseph Y Halpern and Max Kleiman-Weiner. Towards formal definitions of blameworthiness, intention, and moral responsibility. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

[Horty and Belnap, 1995] John F Horty and Nuel Belnap. The deliberative STIT: A study of action, omission, ability, and obligation. *Journal of Philosophical Logic*, 24(6):583–644, 1995.

[Horty and Pacuit, 2017] John Horty and Eric Pacuit. Action types in STIT semantics. *The Review of Symbolic Logic*, pages 1–21, 2017.

[Horty, 2001] John F Horty. *Agency and deontic logic*. Oxford University Press, 2001.

[Kagan, 1991] Shelly Kagan. *The Limits of Morality*. Oxford Ethics Series. Clarendon Press, 1991.

[Mendelson, 2009] Elliott Mendelson. *Introduction to mathematical logic*. CRC press, 2009.

[Naumov and Tao, 2019] Pavel Naumov and Jia Tao. Blameworthiness in strategic games. In *Proceedings of Thirty-third AAAI Conference on Artificial Intelligence (AAAI-19)*, 2019.

[Naumov and Tao, 2020a] Pavel Naumov and Jia Tao. Blameworthiness in security games. In *Proceedings of Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, 2020.

[Naumov and Tao, 2020b] Pavel Naumov and Jia Tao. Duty to warn in strategic games. In Amal El Fallah Seghrouchni, Gita Sukthankar, Bo An, and Neil Yorke-Smith, editors, *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20, Auckland, New Zealand, May 9-13, 2020*, pages 904–912. International Foundation for Autonomous Agents and Multiagent Systems, 2020.

[Naumov and Tao, 2020c] Pavel Naumov and Jia Tao. An epistemic logic of blameworthiness. *Artificial Intelligence*, 283, June 2020. 103269.

[Naumov and Yew, 2021] Pavel Naumov and Rui-Jie Yew. Ethical dilemmas in strategic games. In *Proceedings of Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, 2021.

[Olkhovikov and Wansing, 2018] Grigory K Olkhovikov and Heinrich Wansing. Inference as doxastic agency. part i: The basics of justification STIT logic. *Studia Logica*, pages 1–28, 2018.

[Widerker, 2017] David Widerker. *Moral responsibility and alternative possibilities: Essays on the importance of alternative possibilities*. Routledge, 2017.