Blameworthiness in Games with Imperfect Information

Pavel Naumov
Claremont McKenna College
Claremont, California
pgn2@cornell.edu

Jia Tao
Lafayette College
Easton, Pennsylvania
taoj@lafayette.edu

ABSTRACT

Blameworthiness of an agent or a coalition of agents is often defined in terms of the principle of alternative possibilities: for the coalition to be responsible for an outcome, the outcome must take place and the coalition should have had a strategy to prevent it. In this paper we argue that in the settings with imperfect information, not only should the coalition have had a strategy, but it also should have known that it had a strategy, and it should have known what the strategy was.

The main technical result of the paper is a sound and complete bimodal logic that describes the interplay between knowledge and blameworthiness in strategic games with imperfect information.

KEYWORDS

blameworthiness; know-how; modal logic; axiomatization

1 INTRODUCTION

In this paper we study blameworthiness of agents and their coalitions in multiagent systems. Throughout centuries, blameworthiness, especially in the context of free will and moral responsibility, has been at the focus of philosophical discussions [46]. These discussions continue in the modern time [15, 16, 33, 41, 51]. Frankfurt acknowledges that a dominant role in these discussions has been played by what he calls a *principle of alternate possibilities*: "a person is morally responsible for what he has done only if he could have done otherwise" [17]. As with many general principles, this one has many limitations that Frankfurt discusses; for example, when a person is coerced into doing something. Following the established tradition [51], we refer to this principle as the principle of *alternative* possibilities.

Others refer to an alternative possibility as a *counterfactual* possibility [12, 23]. Halpern and Pearl proposed several versions of a formal definition of causality as a relation between sets of variables that include a counterfactual requirement [23]. Halpern and Kleiman-Weiner used a similar setting to define *degrees* of blameworthiness [25]. Batusov and Soutchanski gave a counterfactual-based definition of causality in situation calculus [6]. Alechina, Halpern, and Logan applied counterfactual definition of causality to team plans [4].

Although the principle of alternative possibilities makes sense in the settings with perfect information, it needs to be adjusted for settings with imperfect information. Indeed, consider a traffic situation depicted in Figure 1. A self-driving truck t and a regular car c are approaching an intersection at which truck t must stop to yield to car c. The truck is experiencing a sudden brake failure and it cannot stop, nor can it slow down at the intersection. The truck turns on flashing lights and sends distress signals to other self-driving cars by radio. The driver of car c can see the flashing lights, but she does not receive the radio signal. She can also observe that the truck does not slow down. The driver of car c has two potential strategies to avoid a collision with the truck: to slow down or to accelerate. The driver understands that one of these two strategies will succeed, but since she does not know the exact speed of the truck, she does not know which of the two strategies will succeed. Suppose that the collision could be avoided if the car accelerates, but the car driver decides to slow down. The vehicles collide. According to the principle of alternative possibilities, the driver of the car is responsible for the collision because she had a strategy to avoid the collision but did not use it.

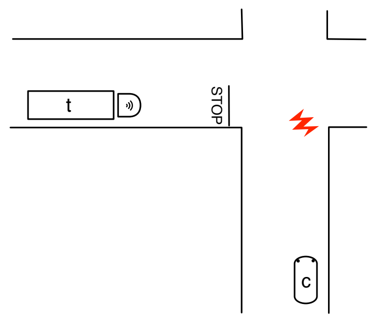

Figure 1: A traffic situation.

It is not likely, however, that a court will find the driver of car c responsible for the accident. For example, US Model Penal Code [29] distinguishes different forms of legal liability as different combinations of "guilty actions" and "guilty mind". The situation in our example falls under strict liability (pure "guilty actions" without an accompanied "guilty mind"). In many situations, strict liability does not lead to legal liability.

In this paper we propose a formal semantics of blameworthiness in strategic games with imperfect information. According to this semantics, an agent (or a coalition of agents) is blamable for φ if φ is true and the agent *knew how* to prevent φ. In our example, since the drive of the car does not know that she must accelerate in order to avoid the collision, she cannot be blamed for the collision. We write this as: $\neg\mathsf{B}_c$("Vehicles collided.").

Now, consider a similar traffic situation in which car c is a self-driving vehicle. The car receives the distress signal from truck t, which contains the truck's exact speed. From this information, car c determines that it can avoid the collision if it accelerates. However, if the car slows down, then the vehicles collide and the self-driving car c is blameable for the collision: B_c("Vehicles collided.").

The main technical result of this paper is a bimodal logical system that describes the interplay between knowledge and blameworthiness of coalitions in strategic games with imperfect information.

2 RELATED LITERATURE

Although the study of responsibility and blameworthiness has a long history in philosophy, the use of formal logical systems to capture these notions is a recent development. Xu proposed a complete logical system for reasoning about responsibility of individual agents in multiagent systems [52]. His approach was extended to coalitions by Broersen, Herzig, and Troquard [11]. The definition of responsibility in these works is different from ours. They assume that an agent or a coalition of agents is responsible for an outcome if the actions that they took unavoidably lead to the outcome. Thus, their definition is not based on the principle of alternative possibilities.

Halpern and Pearl gave several versions of a formal definition of causality between sets of variables using counterfactuals [23]. Lorini and Schwarzentruber [32] observed that a variation of this definition can be captured in STIT logic [8, 26–28, 42]. They said that there is a counterfactual dependence between actions of a coalition C and an outcome φ if φ is true and the complement of the coalition C had no strategy to force φ. In their notations: $\mathsf{CHP}_C\varphi \equiv \varphi \wedge \neg[\mathcal{A}\backslash C]\,\varphi$, where $\mathcal{A}$ is the set of all agents. They also observed that many human emotions (regret, rejoice, disappointment, elation) can be expressed through a combination of the modality CHP and the knowledge modality.

The game-like setting of this paper closely resembles the semantics of Mark Pauly's logic of coalition power [43, 44]. His approach has been widely investigated in the literature [2, 3, 5, 7, 9, 18, 20–22, 34, 45, 48]. Logics of coalition power study modality that express what a coalition *can do*. We modified Mark Pauly's semantics to express what a coalition *could have done* [37]. We axiomatized a logic that combines statements "φ is true" and "coalition C could have prevented φ" into a single modality $\mathsf{B}_C\varphi$.

In this paper we replace "coalition C could have prevented φ" in [37] with "coalition C knew how it could have prevented φ". The distinction between an agent having a strategy, knowing that the strategy exists, and knowing what the strategy is has been studied before. While Jamroga and Ågotnes talked about "knowledge to identify and execute a strategy" [30], Jamroga and van der Hoek discussed "difference between an agent knowing that he has a suitable strategy and knowing the strategy itself" [31]. Van Benthem called such strategies "uniform" [47]. Broersen talked about "knowingly doing" [10], while Broersen, Herzig, and Troquard discussed modality "know they can do" [11]. We used term "executable strategy" [35]. Wang talked about "knowing how" [49, 50]. The properties of know-how as a modality have been previously axiomatized in different settings [1, 14, 35, 36, 38–40, 49, 50].

The axioms of the logical system proposed in this paper are very similar to the axioms in [37] for blameworthiness in games with perfect information and so are the proofs of soundness of these axioms. The most important contribution of this paper is the proof of completeness, in which the construction from [37] is significantly modified to incorporate distributed knowledge. These modifications are discussed in the beginning of Section 8. The increment from [37] to the current paper is similar to the one from original Mark Pauly's logic of coalition power [43, 44] to more recent works on the interplay of knowledge and know-how modalities [1, 14, 35, 36, 38–40].

3 OUTLINE

The paper is organized as follows: Section 4 presents the formal syntax and semantics of our logical system. Section 5 introduces our axioms and compares them to those in the related works. Section 6 gives examples of formal derivations in the proposed logical system. Sections 7 and 8 prove the soundness and the completeness of our system. Section 9 concludes with a discussion of future work.

4 SYNTAX AND SEMANTICS

In this paper we assume a fixed set $\mathcal{A}$ of agents and a fixed set of propositional variables Prop. By a coalition we mean an arbitrary subset of set $\mathcal{A}$.

Definition 4.1. Φ is the minimal set of formulae such that

(1) $p \in \Phi$ for each variable $p \in$ Prop,
(2) $\varphi \to \psi, \neg\varphi \in \Phi$ for all formulae $\varphi, \psi \in \Phi$,
(3) $\mathsf{K}_C\varphi$, $\mathsf{B}_C\varphi \in \Phi$ for each coalition $C \subseteq \mathcal{A}$ and each $\varphi \in \Phi$.

In other words, language Φ is defined by grammar:

$$\varphi := p \mid \neg\varphi \mid \varphi \to \varphi \mid \mathsf{K}_C\varphi \mid \mathsf{B}_C\varphi.$$

Formula $\mathsf{K}_C\varphi$ is read as "coalition C distributively knew before the actions were taken that statement φ would be true" and formula $\mathsf{B}_C\varphi$ as "coalition C is blamable for φ".

Boolean connectives $\vee, \wedge$, and $\leftrightarrow$ as well as constants $\bot$ and $\top$ are defined in the standard way. By formula $\overline{\mathsf{K}}_C\varphi$ we mean $\neg\mathsf{K}_C\neg\varphi$. For the disjunction of multiple formulae, we assume that parentheses are nested to the left. That is, formula $\chi_1 \vee \chi_2 \vee \chi_3$ is a shorthand for $(\chi_1 \vee \chi_2) \vee \chi_3$. As usual, the empty disjunction is defined to be $\bot$. For any two sets X and Y, by X^Y we denote the set of all functions from Y to X.

The formal semantics of modalities K and B is defined in terms of models, which we call *games*. These are one-shot strategic games with imperfect information. We specify the set of actions by all agents, or a *complete action profile*, as a function $\delta \in \Delta^{\mathcal{A}}$ from the set of all agents $\mathcal{A}$ to the set of all actions Δ.

Definition 4.2. A game is a tuple $(I, \{\sim_a\}_{a\in\mathcal{A}}, \Delta, \Omega, P, \pi)$, where

(1) I is a set of "initial states",
(2) $\sim_a$ is an "indistinguishability" equivalence relation on set I,
(3) Δ is a nonempty set of "actions",
(4) Ω is a set of "outcomes",
(5) the set of "plays" P is an arbitrary set of tuples (α, δ, ω) such that $\alpha \in I$, $\delta \in \Delta^{\mathcal{A}}$, and $\omega \in \Omega$,
(6) π is a function that maps Prop into subsets of P.

In the introductory example, the set I has two states *high* and *low*, corresponding to the truck going at a high or low speed. The drive of the regular car c cannot distinguish these two states while these states can be distinguished by a self-driving version of car c. For the sake of simplicity, assume that there are two actions that car c can take: $\Delta = \{$*slow-down*, *speed-up*$\}$ and only two possible outcomes: $\Omega = \{$*collision*, *no collision*$\}$. Vehicles collide if either the truck goes with a low speed and the car decides to slow-down or the truck goes with a high speed and the car decides to accelerate. In our case there is only one agent (car c), so the complete action profile can be described by giving just the action of this agent. We refer to the two complete action profiles in this situation simply as profile *slow-down* and profile *speed-up*.

The list of all possible scenarios (or "plays") is given by the set

$$P = \{(high, speed\text{-}up, collision), (high, slow\text{-}down, no\ collision), \{(low, speed\text{-}up, no\ collision), (low, slow\text{-}down, collision)\}.$$

Note that in our example an initial state and an action profile uniquely determine the outcome. In general, we allow nondeterministic games where this does not have to be true. We also do not require that, for any initial state and any complete action profile, there is at least one outcome. In other words, in certain situations we allow agents to terminate the game without reaching an outcome. This is a more general setting and it minimizes the list of axioms. If one wishes not to consider such games, an additional axiom $\neg\mathsf{B}_C\top$ should be added to the logical system without any major changes in the proof of the completeness.

Whether statement $\mathsf{B}_C\varphi$ is true or false depends not only on the outcome but also on the initial state of the game. Indeed, coalition C might have known how to prevent φ in one initial state but not in the other. For this reason, we assume that all statements are true or false for a particular play of the game. For example, propositional variable p can stand for "car c slowed down and collided with truck t going at a high speed". As a result, function π in the definition above maps p into subsets of P rather than subsets of Ω.

By an action profile of a coalition C we mean an arbitrary function $s \in \Delta^C$ that assigns an action to each member of the coalition. If s_1 and s_2 are action profiles of coalitions C_1 and C_2, respectively, and C is any coalition such that $C \subseteq C_1 \cap C_2$, then we write $s_1 =_C s_2$ to denote that $s_1(a) = s_2(a)$ for each agent $a \in C$.

Next is the key definition of this paper. Its item 5 formally specifies blameworthiness using the principle of alternative possibilities. In order for a coalition to be blamable for φ, not only must φ be true and the coalition should have had a strategy to prevent φ, but this strategy should work in all initial states that the coalition cannot distinguish from the current state. In other words, the coalition should have known the strategy.

Definition 4.3. For any game $(I, \{\sim_a\}_{a\in\mathcal{A}}, \Delta, \Omega, P, \pi)$, any formula $\varphi \in \Phi$, and any play $(\alpha, \delta, \omega) \in P$, the satisfiability relation $(\alpha, \delta, \omega) \Vdash \varphi$ is defined recursively as follows:

(1) $(\alpha, \delta, \omega) \Vdash p$ if $(\alpha, \delta, \omega) \in \pi(p)$, where $p \in \mathsf{Prop}$,
(2) $(\alpha, \delta, \omega) \Vdash \neg\varphi$ if $(\alpha, \delta, \omega) \nVdash \varphi$,
(3) $(\alpha, \delta, \omega) \Vdash \varphi \to \psi$ if $(\alpha, \delta, \omega) \nVdash \varphi$ or $(\alpha, \delta, \omega) \Vdash \psi$,
(4) $(\alpha, \delta, \omega) \Vdash \mathsf{K}_C\varphi$ if $(\alpha', \delta', \omega') \Vdash \varphi$ for each play $(\alpha', \delta', \omega') \in P$ such that $\alpha \sim_C \alpha'$,
(5) $(\alpha, \delta, \omega) \Vdash \mathsf{B}_C\varphi$ if $(\alpha, \delta, \omega) \Vdash \varphi$ and there is an action profile $s \in \Delta^C$ of coalition C such that for each play $(\alpha', \delta', \omega') \in P$, if $\alpha \sim_C \alpha'$ and $s =_C \delta'$, then $(\alpha', \delta', \omega') \nVdash \varphi$.

Since modality K_C represents *a priori* (before the actions) knowledge of coalition C, only the initial states in plays (α, δ, ω) and $(\alpha', \delta', \omega')$ are indistinguishable in item (4) of Definition 4.3.

Note that in part 5 of the above definition we do not assume that coalition C is a minimal one that knew how to prevent the outcome. This is different from the definition of blameworthiness in [24]. Our approach is consistent with how word "blame" is often used in English. For example, the sentence "Millennials being blamed for decline of American cheese" [19] does not imply that no one in the millennial generation likes American cheese.

5 AXIOMS

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

(1) Truth: $\mathsf{K}_C\varphi \to \varphi$ and $\mathsf{B}_C\varphi \to \varphi$,
(2) Distributivity: $\mathsf{K}_C(\varphi \to \psi) \to (\mathsf{K}_C\varphi \to \mathsf{K}_C\psi)$,
(3) Negative Introspection: $\neg\mathsf{K}_C\varphi \to \mathsf{K}_C\neg\mathsf{K}_C\varphi$,
(4) Monotonicity: $\mathsf{K}_C\varphi \to \mathsf{K}_D\varphi$ and $\mathsf{B}_C\varphi \to \mathsf{B}_D\varphi$, where $C \subseteq D$,
(5) None to Blame: $\neg\mathsf{B}_\varnothing\varphi$,
(6) Joint Responsibility: if $C \cap D = \varnothing$, then $\overline{\mathsf{K}}_C\mathsf{B}_C\varphi \wedge \overline{\mathsf{K}}_D\mathsf{B}_D\psi \to (\varphi \vee \psi \to \mathsf{B}_{C\cup D}(\varphi \vee \psi))$,
(7) Blame for Known Cause: $\mathsf{K}_C(\varphi \to \psi) \to (\mathsf{B}_C\psi \to (\varphi \to \mathsf{B}_C\varphi))$,
(8) Knowledge of Fairness: $\mathsf{B}_C\varphi \to \mathsf{K}_C(\varphi \to \mathsf{B}_C\varphi)$.

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

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

We write $X \vdash \varphi$ if formula φ is provable from the theorems of our logical system and an additional set of axioms X using only the Modus Ponens inference rule.

The Truth, the Distributivity, the Negative Introspection, and the Monotonicity axioms for epistemic modality K are the standard S5 axioms from the logic of distributed knowledge. The Truth axiom for blameworthiness modality B states that a coalition could only be blamed for something true. The Monotonicity axiom for the blameworthiness modality states that if a part of a coalition is blamable for something, then the whole coalition is also blamable for the same thing. The None to Blame axiom says that an empty coalition can be blamed for nothing.

The remaining three axioms describe the interplay between knowledge and blameworthiness modalities.

The Joint Responsibility axiom says that if a coalition C cannot exclude a possibility of being blamable for φ, a coalition D cannot exclude a possibility of being blamable for ψ, and the disjunction $\varphi \vee \psi$ is true, then the joint coalition $C \cup D$ is blamable for the disjunction. This axiom resembles Xu's axiom for the independence of individual agents [52],

$$\overline{\mathsf{N}}\mathsf{B}_{a_1}\varphi_1 \wedge \dots \wedge \overline{\mathsf{N}}\mathsf{B}_{a_n}\varphi_n \to \overline{\mathsf{N}}(\mathsf{B}_{a_1}\varphi_1 \wedge \dots \wedge \mathsf{B}_{a_n}\varphi_n),$$

where modality $\overline{\mathsf{N}}$ is an abbreviation for $\neg\mathsf{N}\neg$ and formula $\mathsf{N}\varphi$ stands for "formula φ is universally true in the given model". Broersen, Herzig, and Troquard [11] captured the independence of disjoint coalitions C and D in their Lemma 17:

$$\overline{\mathsf{N}}\mathsf{B}_C\varphi \wedge \overline{\mathsf{N}}\mathsf{B}_D\psi \to \overline{\mathsf{N}}(\mathsf{B}_C\varphi \wedge \mathsf{B}_D\psi).$$

In spite of certain similarity, the definition of responsibility used in [52] and [11] does not assume the principle of alternative possibilities. The Joint Responsibility axiom is also similar to Marc Pauly's Cooperation axiom for the logic of coalitional power [43, 44]:

$$\mathsf{S}_C\varphi \wedge \mathsf{S}_D\psi \to \mathsf{S}_{C\cup D}(\varphi \wedge \psi),$$

where coalitions C and D are disjoint and $\mathsf{S}_C\varphi$ stands for "coalition C has a strategy to achieve φ". Finally, The Joint Responsibility axiom in this paper is a generalization of the Joint Responsibility axiom for games with perfect information [37]:

$$\overline{\mathsf{N}}\mathsf{B}_C\varphi \wedge \overline{\mathsf{N}}\mathsf{B}_D\psi \to (\varphi \vee \psi \to \mathsf{B}_{C\cup D}(\varphi \vee \psi)),$$

where coalitions C and D are disjoint.

We proposed the Blame for Cause axiom for the games with perfect information [37]:

$$\mathsf{N}(\varphi \to \psi) \to (\mathsf{B}_C\psi \to (\varphi \to \mathsf{B}_C\varphi)).$$

This axiom is interpreted as "if formula φ universally implies ψ (informally, φ is a *cause* of ψ), then any coalition blamable for ψ should also be blamable for the cause φ as long as φ is actually true." The Blame for *Known* Cause axiom generalizes this principle to the games with imperfect information.

The Knowledge of Fairness axiom also goes back to one of axioms for the games with perfect information. The Fairness axiom

$$\mathsf{B}_C\varphi \to \mathsf{N}(\varphi \to \mathsf{B}_C\varphi)$$

states "if a coalition C is blamed for φ, then it should be blamed for φ whenever φ is true" [37]. The Knowledge of Fairness axiom states that if a coalition C is blamable for φ in an imperfect information game, then it *knows* that it is blamable for φ whenever φ is true.

6 EXAMPLES OF DERIVATIONS

We prove soundness of the axioms of our logical system in the next section. Here we prove several lemmas about our formal system that will be used later in the proof of the completeness. All of these lemmas, stated for modality N instead of modality K originally appeared in [37].

Lemma 6.1. $\vdash \overline{\mathsf{K}}_C\mathsf{B}_C\varphi \to (\varphi \to \mathsf{B}_C\varphi)$.

Proof. Note that $\vdash \mathsf{B}_C\varphi \to \mathsf{K}_C(\varphi \to \mathsf{B}_C\varphi)$ by the Knowledge of Fairness axiom. Thus, $\vdash \neg\mathsf{K}_C(\varphi \to \mathsf{B}_C\varphi) \to \neg\mathsf{B}_C\varphi$, by the law of contrapositive. Then, $\vdash \mathsf{K}_C(\neg\mathsf{K}_C(\varphi \to \mathsf{B}_C\varphi) \to \neg\mathsf{B}_C\varphi)$ by the Necessitation inference rule. Hence, by the Distributivity axiom and the Modus Ponens inference rule,

$$\vdash \mathsf{K}_C\neg\mathsf{K}_C(\varphi \to \mathsf{B}_C\varphi) \to \mathsf{K}_C\neg\mathsf{B}_C\varphi.$$

At the same time, by the Negative Introspection axiom:

$$\vdash \neg\mathsf{K}_C(\varphi \to \mathsf{B}_C\varphi) \to \mathsf{K}_C\neg\mathsf{K}_C(\varphi \to \mathsf{B}_C\varphi).$$

Then, by the laws of propositional reasoning,

$$\vdash \neg\mathsf{K}_C(\varphi \to \mathsf{B}_C\varphi) \to \mathsf{K}_C\neg\mathsf{B}_C\varphi.$$

Thus, by the law of contrapositive,

$$\vdash \neg\mathsf{K}_C\neg\mathsf{B}_C\varphi \to \mathsf{K}_C(\varphi \to \mathsf{B}_C\varphi).$$

Since $\mathsf{K}_C(\varphi \to \mathsf{B}_C\varphi) \to (\varphi \to \mathsf{B}_C\varphi)$ is an instance of the Truth axiom, by propositional reasoning,

$$\vdash \neg\mathsf{K}_C\neg\mathsf{B}_C\varphi \to (\varphi \to \mathsf{B}_C\varphi).$$

Therefore, $\vdash \overline{\mathsf{K}}_C\mathsf{B}_C\varphi \to (\varphi \to \mathsf{B}_C\varphi)$ by the definition of $\overline{\mathsf{K}}_C$. □

Lemma 6.2. *If* $\vdash \varphi \leftrightarrow \psi$*, then* $\vdash \mathsf{B}_C\varphi \to \mathsf{B}_C\psi$.

Proof. By the Blame for Known Cause axiom,

$$\vdash \mathsf{K}_C(\psi \to \varphi) \to (\mathsf{B}_C\varphi \to (\psi \to \mathsf{B}_C\psi)).$$

Assumption $\vdash \varphi \leftrightarrow \psi$ implies $\vdash \psi \to \varphi$ by the laws of propositional reasoning. Hence, $\vdash \mathsf{K}_C(\psi \to \varphi)$ by the Necessitation inference rule. Thus, by the Modus Ponens rule, $\vdash \mathsf{B}_C\varphi \to (\psi \to \mathsf{B}_C\psi)$. Then, by the laws of propositional reasoning,

$$\vdash (\mathsf{B}_C\varphi \to \psi) \to (\mathsf{B}_C\varphi \to \mathsf{B}_C\psi). \tag{1}$$

Observe that $\vdash \mathsf{B}_C\varphi \to \varphi$ by the Truth axiom. Also, $\vdash \varphi \leftrightarrow \psi$ by the assumption of the lemma. Then, by the laws of propositional reasoning, $\vdash \mathsf{B}_C\varphi \to \psi$. Therefore, $\vdash \mathsf{B}_C\varphi \to \mathsf{B}_C\psi$ by the Modus Ponens inference rule from statement (1). □

Lemma 6.3. $\varphi \vdash \overline{\mathsf{K}}_C\varphi$.

Proof. By the Truth axioms, $\vdash \mathsf{K}_C\neg\varphi \to \neg\varphi$. Hence, by the law of contrapositive, $\vdash \varphi \to \neg\mathsf{K}_C\neg\varphi$. Thus, $\vdash \varphi \to \overline{\mathsf{K}}_C\varphi$ by the definition of the modality $\overline{\mathsf{K}}_C$. Therefore, $\varphi \vdash \overline{\mathsf{K}}_C\varphi$ by the Modus Ponens inference rule. □

The next lemma generalizes the Joint Responsibility axiom from two coalitions to multiple coalitions.

Lemma 6.4. *For any integer $n \geq 0$ and any pairwise disjoint sets $D_1, \dots, D_n$,*

$$\{\overline{\mathsf{K}}_{D_i}\mathsf{B}_{D_i}\chi_i\}_{i=1}^n, \chi_1 \vee \dots \vee \chi_n \vdash \mathsf{B}_{D_1\cup\dots\cup D_n}(\chi_1 \vee \dots \vee \chi_n).$$

Proof. We prove the lemma by induction on n. If $n = 0$, then disjunction $\chi_1 \vee \dots \vee \chi_n$ is Boolean constant false $\bot$. Hence, the statement of the lemma, $\bot \vdash \mathsf{B}_\varnothing\bot$, is provable in the propositional logic.

Next, assume that $n = 1$. Then, from Lemma 6.1 using Modus Ponens rule twice, we get $\overline{\mathsf{K}}_{D_1}\mathsf{B}_{D_1}\chi_1, \chi_1 \vdash \mathsf{B}_{D_1}\chi_1$.

Assume now that $n \geq 2$. By the Joint Responsibility axiom and the Modus Ponens inference rule,

$$\overline{\mathsf{K}}_{D_1\cup\dots\cup D_{n-1}}\mathsf{B}_{D_1\cup\dots\cup D_{n-1}}(\chi_1 \vee \dots \vee \chi_{n-1}), \overline{\mathsf{K}}_{D_n}\mathsf{B}_{D_n}\chi_n,$$
$$\chi_1 \vee \dots \vee \chi_{n-1} \vee \chi_n \vdash \mathsf{B}_{D_1\cup\dots\cup D_{n-1}\cup D_n}(\chi_1 \vee \dots \vee \chi_{n-1} \vee \chi_n).$$

Hence, by Lemma 6.3,

$$\mathsf{B}_{D_1\cup\dots\cup D_{n-1}}(\chi_1 \vee \dots \vee \chi_{n-1}), \overline{\mathsf{K}}_{D_n}\mathsf{B}_{D_n}\chi_n, \chi_1 \vee \dots \vee \chi_{n-1} \vee \chi_n$$
$$\vdash \mathsf{B}_{D_1\cup\dots\cup D_{n-1}\cup D_n}(\chi_1 \vee \dots \vee \chi_{n-1} \vee \chi_n).$$

At the same time, by the induction hypothesis,

$$\{\overline{\mathsf{K}}_{D_i}\mathsf{B}_{D_i}\chi_i\}_{i=1}^{n-1}, \chi_1 \vee \dots \vee \chi_{n-1} \vdash \mathsf{B}_{D_1\cup\dots\cup D_{n-1}}(\chi_1 \vee \dots \vee \chi_{n-1}).$$

Thus,

$$\{\overline{\mathsf{K}}_{D_i}\mathsf{B}_{D_i}\chi_i\}_{i=1}^{n}, \chi_1 \vee \dots \vee \chi_{n-1}, \chi_1 \vee \dots \vee \chi_{n-1} \vee \chi_n$$
$$\vdash \mathsf{B}_{D_1\cup\dots\cup D_{n-1}\cup D_n}(\chi_1 \vee \dots \vee \chi_{n-1} \vee \chi_n).$$

Note that $\chi_1 \vee \dots \vee \chi_{n-1} \vdash \chi_1 \vee \dots \vee \chi_{n-1} \vee \chi_n$ is provable in the propositional logic. Thus,

$$\{\overline{\mathsf{K}}_{D_i}\mathsf{B}_{D_i}\chi_i\}_{i=1}^{n}, \chi_1 \vee \dots \vee \chi_{n-1}$$
$$\vdash \mathsf{B}_{D_1\cup\dots\cup D_{n-1}\cup D_n}(\chi_1 \vee \dots \vee \chi_{n-1} \vee \chi_n). \tag{2}$$

Similarly, by the Joint Responsibility axiom and the Modus Ponens inference rule,

$$\overline{\mathsf{K}}_{D_1}\mathsf{B}_{D_1}\chi_1, \overline{\mathsf{K}}_{D_2\cup\dots\cup D_n}\mathsf{B}_{D_2\cup\dots\cup D_n}(\chi_2 \vee \dots \vee \chi_n),$$
$$\chi_1 \vee (\chi_2 \vee \dots \vee \chi_n) \vdash \mathsf{B}_{D_1\cup\dots\cup D_{n-1}\cup D_n}(\chi_1 \vee (\chi_2 \vee \dots \vee \chi_n)).$$

Because formula $\chi_1 \vee (\chi_2 \vee \dots \vee \chi_n) \leftrightarrow \chi_1 \vee \chi_2 \vee \dots \vee \chi_n$ is provable in the propositional logic, by Lemma 6.2,

$$\overline{\mathsf{K}}_{D_1}\mathsf{B}_{D_1}\chi_1, \overline{\mathsf{K}}_{D_2\cup\dots\cup D_n}\mathsf{B}_{D_2\cup\dots\cup D_n}(\chi_2 \vee \dots \vee \chi_n),$$
$$\chi_1 \vee \chi_2 \vee \dots \vee \chi_n \vdash \mathsf{B}_{D_1\cup\dots\cup D_{n-1}\cup D_n}(\chi_1 \vee \chi_2 \vee \dots \vee \chi_n).$$

Hence, by Lemma 6.3,

$$\overline{\mathsf{K}}_{D_1}\mathsf{B}_{D_1}\chi_1, \mathsf{B}_{D_2\cup\dots\cup D_n}(\chi_2\vee\dots\vee\chi_n), \chi_1\vee\chi_2\vee\dots\vee\chi_n \vdash \mathsf{B}_{D_1\cup\dots\cup D_{n-1}\cup D_n}(\chi_1\vee\chi_2\vee\dots\vee\chi_n).$$

At the same time, by the induction hypothesis,

$$\{\overline{\mathsf{K}}_{D_i}\mathsf{B}_{D_i}\chi_i\}_{i=2}^n, \chi_2\vee\dots\vee\chi_n \vdash \mathsf{B}_{D_2\cup\dots\cup D_n}(\chi_2\vee\dots\vee\chi_n).$$

Thus,

$$\{\overline{\mathsf{K}}_{D_i}\mathsf{B}_{D_i}\chi_i\}_{i=1}^n, \chi_2\vee\dots\vee\chi_n, \chi_1\vee\chi_2\vee\dots\vee\chi_n \vdash \mathsf{B}_{D_1\cup D_2\cup\dots\cup D_n}(\chi_1\vee\chi_2\vee\dots\vee\chi_n).$$

Note that $\chi_2\vee\dots\vee\chi_n \vdash \chi_1\vee\dots\vee\chi_{n-1}\vee\chi_n$ is provable in the propositional logic. Thus,

$$\{\overline{\mathsf{K}}_{D_i}\mathsf{B}_{D_i}\chi_i\}_{i=1}^n, \chi_2\vee\dots\vee\chi_n \vdash \mathsf{B}_{D_1\cup\dots\cup D_{n-1}\cup D_n}(\chi_1\vee\chi_2\vee\dots\vee\chi_n). \tag{3}$$

Finally, note that the following statement is provable in the propositional logic for $n\geq 2$,

$$\vdash \chi_1\vee\dots\vee\chi_n \to (\chi_1\vee\dots\vee\chi_{n-1})\vee(\chi_2\vee\dots\vee\chi_n).$$

Therefore, from statement (2) and statement (3)

$$\{\overline{\mathsf{K}}_{D_i}\mathsf{B}_{D_i}\chi_i\}_{i=1}^n, \chi_1\vee\dots\vee\chi_n \vdash \mathsf{B}_{D_1\cup\dots\cup D_n}(\chi_1\vee\dots\vee\chi_n).$$

by the laws of propositional reasoning. □

Lemma 6.5. *If $\varphi_1,\dots,\varphi_n\vdash\psi$, then $\mathsf{K}_C\varphi_1,\dots,\mathsf{K}_C\varphi_n\vdash\mathsf{K}_C\psi$.*

Proof. By the deduction lemma applied n times, assumption $\varphi_1,\dots,\varphi_n\vdash\psi$ implies that $\vdash\varphi_1\to(\varphi_2\to\dots(\varphi_n\to\psi)\dots)$. Thus, by the Necessitation inference rule,

$$\vdash \mathsf{K}_C(\varphi_1\to(\varphi_2\to\dots(\varphi_n\to\psi)\dots)).$$

Hence, by the Distributivity axiom and the Modus Ponens rule,

$$\vdash \mathsf{K}_C\varphi_1\to\mathsf{K}_C(\varphi_2\to\dots(\varphi_n\to\psi)\dots).$$

Then, again by the Modus Ponens rule,

$$\mathsf{K}_C\varphi_1\vdash\mathsf{K}_C(\varphi_2\to\dots(\varphi_n\to\psi)\dots).$$

Therefore, $\mathsf{K}_C\varphi_1,\dots,\mathsf{K}_C\varphi_n\vdash\mathsf{K}_C\psi$ by applying the previous steps $(n-1)$ more times. □

The following lemma states a well-known principle in epistemic logic. The proof of this principle can be found, for example, in [40].

Lemma 6.6 (Positive Introspection). $\vdash\mathsf{K}_C\varphi\to\mathsf{K}_C\mathsf{K}_C\varphi$. □

Our last example rephrases Lemma 6.4 into the form which is used in the proof of the completeness.

Lemma 6.7. *For any $n\geq 0$ and any disjoint sets $D_1,\dots,D_n\subseteq C$,*

$$\{\overline{\mathsf{K}}_{D_i}\mathsf{B}_{D_i}\chi_i\}_{i=1}^n, \mathsf{K}_C(\varphi\to\chi_1\vee\dots\vee\chi_n)\vdash\mathsf{K}_C(\varphi\to\mathsf{B}_C\varphi).$$

Proof. By Lemma 6.4,

$$\{\overline{\mathsf{K}}_{D_i}\mathsf{B}_{D_i}\chi_i\}_{i=1}^n, \chi_1\vee\dots\vee\chi_n \vdash \mathsf{B}_{D_1\cup\dots\cup D_n}(\chi_1\vee\dots\vee\chi_n).$$

Hence, by the Monotonicity axiom,

$$\{\overline{\mathsf{K}}_{D_i}\mathsf{B}_{D_i}\chi_i\}_{i=1}^n, \chi_1\vee\dots\vee\chi_n \vdash \mathsf{B}_C(\chi_1\vee\dots\vee\chi_n).$$

Thus, by the Modus Ponens inference rule

$$\{\overline{\mathsf{K}}_{D_i}\mathsf{B}_{D_i}\chi_i\}_{i=1}^n, \varphi, \varphi\to\chi_1\vee\dots\vee\chi_n \vdash \mathsf{B}_C(\chi_1\vee\dots\vee\chi_n).$$

By the Truth axiom and the Modus Ponens inference rule,

$$\{\overline{\mathsf{K}}_{D_i}\mathsf{B}_{D_i}\chi_i\}_{i=1}^n, \varphi, \mathsf{K}_C(\varphi\to\chi_1\vee\dots\vee\chi_n) \vdash \mathsf{B}_C(\chi_1\vee\dots\vee\chi_n).$$

The following formula is an instance of the Blame for Known Cause axiom $\mathsf{K}_C(\varphi\to\chi_1\vee\dots\vee\chi_n)\to(\mathsf{B}_C(\chi_1\vee\dots\vee\chi_n)\to(\varphi\to\mathsf{B}_C\varphi))$. Hence, by the Modus Ponens inference rule applied twice,

$$\{\overline{\mathsf{K}}_{D_i}\mathsf{B}_{D_i}\chi_i\}_{i=1}^n, \varphi, \mathsf{K}_C(\varphi\to\chi_1\vee\dots\vee\chi_n) \vdash \varphi\to\mathsf{B}_C\varphi.$$

By the Modus Ponens inference rule,

$$\{\overline{\mathsf{K}}_{D_i}\mathsf{B}_{D_i}\chi_i\}_{i=1}^n, \varphi, \mathsf{K}_C(\varphi\to\chi_1\vee\dots\vee\chi_n) \vdash \mathsf{B}_C\varphi.$$

By the deduction lemma,

$$\{\overline{\mathsf{K}}_{D_i}\mathsf{B}_{D_i}\chi_i\}_{i=1}^n, \mathsf{K}_C(\varphi\to\chi_1\vee\dots\vee\chi_n) \vdash \varphi\to\mathsf{B}_C\varphi.$$

By Lemma 6.5,

$$\{\mathsf{K}_C\overline{\mathsf{K}}_{D_i}\mathsf{B}_{D_i}\chi_i\}_{i=1}^n, \mathsf{K}_C\mathsf{K}_C(\varphi\to\chi_1\vee\dots\vee\chi_n) \vdash \mathsf{K}_C(\varphi\to\mathsf{B}_C\varphi).$$

By the Monotonicity axiom, the Modus Ponens inference rule, and the assumption $D_1,\dots,D_n\subseteq C$,

$$\{\mathsf{K}_{D_i}\overline{\mathsf{K}}_{D_i}\mathsf{B}_{D_i}\chi_i\}_{i=1}^n, \mathsf{K}_C\mathsf{K}_C(\varphi\to\chi_1\vee\dots\vee\chi_n) \vdash \mathsf{K}_C(\varphi\to\mathsf{B}_C\varphi).$$

By the definition of modality $\overline{\mathsf{K}}$, the Negative Introspection axiom, and the Modus Ponens inference rule,

$$\{\overline{\mathsf{K}}_{D_i}\mathsf{B}_{D_i}\chi_i\}_{i=1}^n, \mathsf{K}_C\mathsf{K}_C(\varphi\to\chi_1\vee\dots\vee\chi_n) \vdash \mathsf{K}_C(\varphi\to\mathsf{B}_C\varphi).$$

Therefore, by Lemma 6.6 and the Modus Ponens inference rule, the statement of the lemma follows. □

7 SOUNDNESS

The epistemic part of the Truth axiom as well as the Distribitivity, the Negative Introspection, and the Monotonicity axioms are the standard axioms of epistemic logic S5 for distributed knowledge. Their soundness follows from the assumption that $\sim_a$ is an equivalence relation in the standard way [13]. The soundness of the blameworthiness part of the Truth axiom and of the Monotonicity axiom immediately follows from Definition 4.3. In this section, we prove the soundness of each of the remaining axioms as a separate lemma. In these lemmas, $C,D\subseteq\mathcal{A}$ are coalitions, $\varphi,\psi\in\Phi$ are formulae, and $(\alpha,\delta,\omega)\in P$ is a play of a game $(I,\{\sim_a\}_{a\in\mathcal{A}},\Delta,\Omega,P,\pi)$.

Lemma 7.1. $(\alpha,\delta,\omega)\nVdash\mathsf{B}_\varnothing\varphi$.

Proof. Assume that $(\alpha,\delta,\omega)\Vdash\mathsf{B}_\varnothing\varphi$. Hence, by Definition 4.3, we have $(\alpha,\delta,\omega)\Vdash\varphi$ and there is an action profile $s\in\Delta^\varnothing$ such that for each play $(\alpha',\delta',\omega')\in P$, if $\alpha\sim_\varnothing\alpha'$ and $s=_\varnothing\delta'$, then $(\alpha',\delta',\omega')\nVdash\varphi$.

Let $\alpha'=\alpha$, $\delta'=\delta$, and $\omega'=\omega$. Since $\alpha\sim_\varnothing\alpha'$ and $s=_\varnothing\delta'$, by the choice of action profile s we have $(\alpha',\delta',\omega')\nVdash\varphi$. Then, $(\alpha,\delta,\omega)\nVdash\varphi$, which leads to a contradiction. □

Lemma 7.2. *If $C\cap D=\varnothing$, $(\alpha,\delta,\omega)\Vdash\overline{\mathsf{K}}_C\mathsf{B}_C\varphi$, $(\alpha,\delta,\omega)\Vdash\overline{\mathsf{K}}_D\mathsf{B}_D\psi$, and $(\alpha,\delta,\omega)\Vdash\varphi\vee\psi$, then $(\alpha,\delta,\omega)\Vdash\mathsf{B}_{C\cup D}(\varphi\vee\psi)$.*

Proof. Suppose that $(\alpha,\delta,\omega)\Vdash\overline{\mathsf{K}}_C\mathsf{B}_C\varphi$ and $(\alpha,\delta,\omega)\Vdash\overline{\mathsf{K}}_D\mathsf{B}_D\psi$. Hence, by Definition 4.3 and the definition of modality $\overline{\mathsf{K}}$, there are plays $(\alpha_1,\delta_1,\omega_1)\in P$ and $(\alpha_2,\delta_2,\omega_2)\in P$ such that $\alpha\sim_C\alpha_1$, $\alpha\sim_D\alpha_2$, $(\alpha_1,\delta_1,\omega_1)\Vdash\mathsf{B}_C\varphi$ and $(\alpha_2,\delta_2,\omega_2)\Vdash\mathsf{B}_D\psi$.

Statement $(\alpha_1,\delta_1,\omega_1)\Vdash\mathsf{B}_C\varphi$, by Definition 4.3, implies that there is a profile $s_1\in\Delta^C$ such that for each play $(\alpha',\delta',\omega')\in P$, if $\alpha_1\sim_C\alpha'$ and $s_1=_C\delta'$, then $(\alpha',\delta',\omega')\nVdash\varphi$.

Similarly, statement $(\alpha_2, \delta_2, \omega_2) \Vdash \mathsf{B}_D\psi$, by Definition 4.3, implies that there is an action profile $s_2 \in \Delta^D$ such that for each play $(\alpha', \delta', \omega') \in P$, if $\alpha_2 \sim_D \alpha'$ and $s_2 =_D \delta'$, then $(\alpha', \delta', \omega') \nVdash \psi$.

Consider an action profile s of coalition $C \cup D$ such that

$$s(a) = \begin{cases} s_1(a), & \text{if } a \in C, \\ s_2(a), & \text{if } a \in D. \end{cases}$$

The action profile s is well-defined because sets C and D are disjoint by the assumption of the lemma.

The choice of action profiles s_1, s_2, and s implies that for each play $(\alpha', \delta', \omega') \in P$, if $\alpha \sim_{C\cup D} \alpha'$ and $s =_{C\cup D} \delta'$, then $(\alpha', \delta', \omega') \nVdash \varphi$ and $(\alpha', \delta', \omega') \nVdash \psi$. Thus, if $\alpha \sim_{C\cup D} \alpha'$ and $s =_{C\cup D} \delta'$, then $(\alpha', \delta', \omega') \nVdash \varphi \vee \psi$, for each play $(\alpha', \delta', \omega') \in P$. Therefore, $(\alpha, \delta, \omega) \Vdash \mathsf{B}_{C\cup D}(\varphi \vee \psi)$ by Definition 4.3 and the assumption $(\alpha, \delta, \omega) \Vdash \varphi \vee \psi$ of the lemma. □

Lemma 7.3. *If $(\alpha, \delta, \omega) \Vdash \mathsf{K}_C(\varphi \to \psi)$, $(\alpha, \delta, \omega) \Vdash \mathsf{B}_C\psi$, and $(\alpha, \delta, \omega) \Vdash \varphi$, then $(\alpha, \delta, \omega) \Vdash \mathsf{B}_C\varphi$.*

Proof. By Definition 4.3, assumption $(\alpha, \delta, \omega) \Vdash \mathsf{K}_C(\varphi \to \psi)$ implies that for each play $(\alpha', \delta', \omega') \in P$ of the game if $\alpha \sim_C \alpha'$, then $(\alpha', \delta', \omega') \Vdash \varphi \to \psi$.

By Definition 4.3, assumption $(\alpha, \delta, \omega) \Vdash \mathsf{B}_C\psi$ implies that there is an action profile $s \in \Delta^C$ such that for each play $(\alpha', \delta', \omega') \in P$, if $\alpha \sim_C \alpha'$ and $s =_C \delta'$, then $(\alpha', \delta', \omega') \nVdash \psi$.

Hence, for each play $(\alpha', \delta', \omega') \in P$, if $\alpha \sim_C \alpha'$ and $s =_C \delta'$, then $(\alpha', \delta', \omega') \nVdash \varphi$. Therefore, $(\alpha, \delta, \omega) \Vdash \mathsf{B}_C\varphi$ by Definition 4.3 and the assumption $(\alpha, \delta, \omega) \Vdash \varphi$ of the lemma. □

Lemma 7.4. *If $(\alpha, \delta, \omega) \Vdash \mathsf{B}_C\varphi$, then $(\alpha, \delta, \omega) \Vdash \mathsf{K}_C(\varphi \to \mathsf{B}_C\varphi)$.*

Proof. By Definition 4.3, assumption $(\alpha, \delta, \omega) \Vdash \mathsf{B}_C\varphi$ implies that there is an action profile $s \in \Delta^C$ such that for each play $(\alpha', \delta', \omega') \in P$, if $\alpha \sim_C \alpha'$ and $s =_C \delta'$, then $(\alpha', \delta', \omega') \nVdash \varphi$.

Let $(\alpha', \delta', \omega') \in P$ be a play where $\alpha \sim_C \alpha'$ and $(\alpha', \delta', \omega') \Vdash \varphi$. By Definition 4.3, it suffices to show that $(\alpha', \delta', \omega') \Vdash \mathsf{B}_C\varphi$.

Consider any play $(\alpha'', \delta'', \omega'') \in P$ such that $\alpha' \sim_C \alpha''$ and $s =_C \delta''$. Then, since $\sim_C$ is an equivalence relation, assumptions $\alpha \sim_C \alpha'$ and $\alpha' \sim_C \alpha''$ imply $\alpha \sim_C \alpha''$. Thus, $(\alpha'', \delta'', \omega'') \nVdash \varphi$ by the choice of action profile s. Therefore, $(\alpha', \delta', \omega') \Vdash \mathsf{B}_C\varphi$ by Definition 4.3 and the assumption $(\alpha', \delta', \omega') \Vdash \varphi$. □

8 COMPLETENESS

In this section we prove the completeness of our logical system. The standard completeness proof for epistemic logic of individual knowledge defines states as maximal consistent sets. Similarly, in [37], we define outcomes of the game as maximal consistent sets. In the case of the epistemic logic of distributed knowledge, two states are usually defined to be indistinguishable by an agent a if these two states have the same K_a formulae. Unfortunately, this approach does not work for distributed knowledge. Indeed, two maximal consistent sets that have the same K_a and K_b formulae might have different $\mathsf{K}_{a,b}$ formulae. Such two states would be indistinguishable to agent a and agent b, however, the distributed knowledge of agents a and b in these states will be different. This situation is inconsistent with Definition 4.3. To solve this problem we define outcomes not as maximal consistent sets of formulae, but as nodes of a tree. This approach has been previously used to prove completeness of several logics for know-how modality [35, 36, 38–40].

We start the proof of the completeness by defining the canonical game $G(X_0) = (I, \{\sim_a\}_{a\in\mathcal{A}}, \Delta, \Omega, P, \pi)$ for each maximal consistent set of formulae X_0.

Definition 8.1. The set of outcomes Ω consists of all finite sequences $X_0, C_1, X_1, C_2, \dots, C_n, X_n$, such that

(1) $n \geq 0$,
(2) X_i is a maximal consistent subset of Φ for each $i \geq 1$,
(3) C_i is a coalition for each $i \geq 1$,
(4) $\{\varphi \mid \mathsf{K}_{C_i}\varphi \in X_{i-1}\} \subseteq X_i$ for each $i \geq 1$.

For any sequence $s = x_1, \dots, x_n$ and any element y, by $s :: y$ we mean the sequence $x_1, \dots, x_n, y$. By $hd(s)$ we mean element x_n. We define a tree structure on the set of outcomes Ω by saying that outcome (node) $\omega = X_0, C_1, X_1, C_2, \dots, C_n, X_n$ and outcome (node) $\omega :: C_{n+1} :: X_{n+1}$ are connected by an undirected edge labeled with all agents in coalition C_{n+1}, see Figure 2.

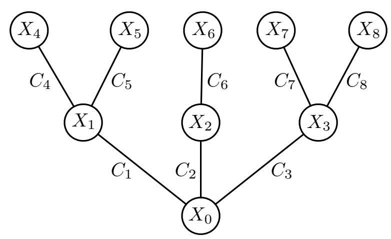

Figure 2: A fragment of tree.

Definition 8.2. For any outcomes $\omega, \omega' \in \Omega$ and any agent $a \in \mathcal{A}$, let $\omega \sim_a \omega'$ if all edges along the unique path between ω and ω' are labeled with agent a.

Lemma 8.3. *Relation $\sim_a$ is an equivalence relation on set Ω.* □

Lemma 8.4. *$\mathsf{K}_D\varphi \in X_n$ iff $\mathsf{K}_D\varphi \in X_{n+1}$ for any formula $\varphi \in \Phi$, any $n \geq 0$, and any outcome $X_0, C_1, X_1, C_2, \dots, X_n, C_{n+1}, X_{n+1} \in \Omega$, and any coalition $D \subseteq C_{n+1}$.*

Proof. If $\mathsf{K}_D\varphi \in X_n$, then $X_n \vdash \mathsf{K}_D\mathsf{K}_D\varphi$ by Lemma 6.6. Hence, $X_n \vdash \mathsf{K}_{C_{n+1}}\mathsf{K}_D\varphi$ by the Monotonicity axiom and the assumption $D \subseteq C_{n+1}$. Thus, $\mathsf{K}_{C_{n+1}}\mathsf{K}_D\varphi \in X_n$ by the maximality of set X_n. Therefore, $\mathsf{K}_D\varphi \in X_{n+1}$ by Definition 8.1.

Suppose that $\mathsf{K}_D\varphi \notin X_n$. Hence, $\neg\mathsf{K}_D\varphi \in X_n$ by the maximality of set X_n. Thus, $X_n \vdash \mathsf{K}_D\neg\mathsf{K}_D\varphi$ by the Negative Introspection axiom. Hence, $X_n \vdash \mathsf{K}_{C_{n+1}}\neg\mathsf{K}_D\varphi$ by the Monotonicity axiom and the assumption $D \subseteq C_{n+1}$. Then, $\mathsf{K}_{C_{n+1}}\neg\mathsf{K}_D\varphi \in X_n$ by the maximality of set X_n. Thus, $\neg\mathsf{K}_D\varphi \in X_{n+1}$ by Definition 8.1. Therefore, $\mathsf{K}_D\varphi \notin X_{n+1}$ because set X_{n+1} is consistent. □

Lemma 8.5. *If $\omega \sim_C \omega'$, then $\mathsf{K}_C\varphi \in hd(\omega)$ iff $\mathsf{K}_C\varphi \in hd(\omega')$.*

Proof. If $\omega \sim_C \omega'$, then each edge along the unique path between nodes ω and ω' is labeled with all agents in coalition C.

We prove the lemma by induction on the length of the unique path between nodes ω and ω'. In the base case, $\omega = \omega'$. Thus, $\mathsf{K}_C\varphi \in hd(\omega)$ iff $\mathsf{K}_C\varphi \in hd(\omega')$. The induction step follows from Lemma 8.4. □

Lemma 8.6. *If $\omega \sim_C \omega'$ and $\mathsf{K}_C\varphi \in hd(\omega)$, then $\varphi \in hd(\omega')$.*

Proof. By Lemma 8.5, assumptions $\omega \sim_C \omega'$ and $\mathsf{K}_C\varphi \in hd(\omega)$ imply that $\mathsf{K}_C\varphi \in hd(\omega')$. Thus, $hd(\omega') \vdash \varphi$ by the Truth axiom and the Modus Ponens inference rule. Therefore, $\varphi \in hd(\omega')$ because set $hd(\omega')$ is maximal. □

The set of the initial states I of the canonical game is the set of all equivalence classes of Ω with respect to relation $\sim_{\mathcal{A}}$.

Definition 8.7. $I = \Omega/\sim_{\mathcal{A}}$.

LEMMA 8.8. *Relation $\sim_C$ is well-defined on set I.*

PROOF. Suppose that $\omega_1 \sim_C \omega_2$. Consider any outcomes ω_1' and ω_2' such that $\omega_1 \sim_{\mathcal{A}} \omega_1'$ and $\omega_2 \sim_{\mathcal{A}} \omega_2'$. It suffices to prove that $\omega_1' \sim_C \omega_2'$.

By Definition 8.2 and Lemma 8.3, assumption $\omega_1 \sim_{\mathcal{A}} \omega_1'$ implies that each edges along the unique path between nodes ω_1' and ω_1 is labeled with all agents in set $\mathcal{A}$. Also, assumption $\omega_1 \sim_C \omega_2$ implies that each edge along the unique path between nodes ω_1 and ω_2 is labeled with all agents in coalition C. Finally, assumption $\omega_2 \sim_{\mathcal{A}} \omega_2'$ implies that each edges along the unique path between nodes ω_2 and ω_2' is labeled with all agents in set $\mathcal{A}$. Hence, each edge along the unique path between nodes ω_1' and ω_2' is labeled with all agents in coalition C. Therefore, $\omega_1' \sim_C \omega_2'$ by Definition 8.2. □

LEMMA 8.9. *$\alpha \sim_C \alpha'$ iff $\omega \sim_C \omega'$, for any initial states $\alpha, \alpha' \in I$, any outcomes $\omega \in \alpha$ and $\omega' \in \alpha'$, and any coalition $C \subseteq \mathcal{A}$.* □

Intuitively, the canonical game consists in agents "vetoing" formulae. The domain of choices of the game consists of all formulae in set Φ. To veto a formula ψ, an agent must choose action ψ. The mechanism guarantees that if $\overline{\mathsf{K}}_C\mathsf{B}_C\psi \in hd(\omega)$ and all agents in the coalition C veto formula ψ, then $\neg\psi \in hd(\omega)$.

Definition 8.10. The domain of actions Δ is set Φ.

Definition 8.11. The set $P \subseteq I \times \Delta^{\mathcal{A}} \times \Omega$ consists of all triples (α, δ, ω) such that $\omega \in \alpha$ and for any formula $\overline{\mathsf{K}}_C\mathsf{B}_C\psi \in hd(\omega)$, if $\delta(a) = \psi$ for each agent $a \in C$, then $\neg\psi \in hd(\omega)$.

Definition 8.12. $\pi(p) = \{(\alpha, \delta, \omega) \in P \mid p \in hd(\omega)\}$.

This concludes the definition of the canonical game $G(X_0)$. We state and prove the completeness later in this section as Theorem 8.17. The four lemmas before are are auxiliary results that will be used in the proof of the completeness.

LEMMA 8.13. *For any play $(\alpha, \delta, \omega) \in P$ of game $G(X_0)$, any action profile $s \in \Delta^C$, and any formula $\neg(\varphi \to \mathsf{B}_C\varphi) \in hd(\omega)$, there is a play $(\alpha', \delta', \omega') \in P$ such that $\alpha \sim_C \alpha'$, $s =_C \delta'$, and $\varphi \in hd(\omega')$.*

PROOF. Consider the following set of formulae:

$$\begin{aligned} X \;=\; & \{\varphi\} \cup \{\psi \mid \mathsf{K}_C\psi \in hd(\omega)\} \\ & \cup \{\neg\chi \mid \overline{\mathsf{K}}_D\mathsf{B}_D\chi \in hd(\omega), D \subseteq C, \forall a \in D(s(a) = \chi)\}. \end{aligned}$$

CLAIM 1. *Set X is consistent.*

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

$$\begin{aligned} \text{formulae} \quad & \mathsf{K}_C\psi_1, \dots, \mathsf{K}_C\psi_m \in hd(\omega), && (4)\\ \text{and formulae} \quad & \overline{\mathsf{K}}_D\mathsf{B}_{D_1}\chi_1, \dots, \overline{\mathsf{K}}_D\mathsf{B}_{D_n}\chi_n \in hd(\omega), && (5)\\ \text{such that} \quad & D_1, \dots, D_n \subseteq C, && (6)\\ & s(a) = \chi_i \text{ for all } i \le n \text{ and all } a \in D_i, && (7)\\ \text{and} \quad & \psi_1, \dots, \psi_m, \neg\chi_1, \dots, \neg\chi_n \vdash \neg\varphi. && (8) \end{aligned}$$

Without loss of generality, we assume that formulae $\chi_1, \dots, \chi_n$ are distinct. Thus, assumption (7) implies that sets $D_1, \dots, D_n$ are pairwise disjoint. By propositional reasoning, assumption (8) implies

$$\psi_1, \dots, \psi_m \vdash \varphi \to \chi_1 \vee \dots \vee \chi_n.$$

Thus, by Lemma 6.5,

$$\mathsf{K}_C\psi_1, \dots, \mathsf{K}_C\psi_m \vdash \mathsf{K}_C(\varphi \to \chi_1 \vee \dots \vee \chi_n).$$

Hence, $hd(\omega) \vdash \mathsf{K}_C(\varphi \to \chi_1 \vee \dots \vee \chi_n)$ by assumption (4). Thus, $hd(\omega) \vdash \mathsf{K}_C(\varphi \to \mathsf{B}_C\varphi)$ by Lemma 6.7, assumption (5), and the assumption that sets $D_1, \dots, D_n$ are pairwise disjoint. Hence, by the Truth axiom, $hd(\omega) \vdash \varphi \to \mathsf{B}_C\varphi$, which contradicts the assumption $\neg(\varphi \to \mathsf{B}_C\varphi) \in hd(\omega)$ of the lemma because set $hd(\omega)$ is consistent. Therefore, set X is consistent. ⊠

Let X' be any maximal consistent extension of set X and ω' be the sequence $\omega :: C :: X'$. Note that $\omega' \in \Omega$ by Definition 8.1 and the choice of sets X and X'. Also $\varphi \in X \subseteq hd(\omega')$ by the choice of sets X and X'.

Let initial state α' be the equivalence class of outcome ω' with respect to the equivalence relation $\sim_{\mathcal{A}}$. Note that $\omega \sim_C \omega'$ by Definition 8.1 and the choice of sequence ω'. Therefore, $\alpha \sim_C \alpha'$ by Lemma 8.9.

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

$$\delta'(a) = \begin{cases} s(a), & \text{if } a \in C,\\ \bot, & \text{otherwise.} \end{cases} \tag{9}$$

Then, $s =_C \delta'$.

CLAIM 2. $(\alpha', \delta', \omega') \in P$.

PROOF OF CLAIM. First, note that $\omega' \in \alpha'$ because state α' is the equivalence class of outcome ω'. Next, consider any formula $\overline{\mathsf{K}}_D\mathsf{B}_D\chi \in hd(\omega')$ such that $\delta'(a) = \chi$ for each $a \in D$. By Definition 8.11, it suffices to show that $\neg\chi \in hd(\omega')$.

Case I: $D \subseteq C$. Thus, $s(a) = \chi$ for each $a \in D$ by equation (9) and the assumption that $\delta'(a) = \chi$ for each $a \in D$.

Suppose that $\neg\chi \notin hd(\omega')$. Then, $\neg\chi \notin X$ because $X \subseteq X' = hd(\omega')$ by the choice of X' and ω'. Thus, $\overline{\mathsf{K}}_D\mathsf{B}_D\chi \notin hd(\omega)$ by the definition of set X and because $s(a) = \chi$ for each $a \in D$. Hence, $\mathsf{K}_D\neg\mathsf{B}_D\chi \in hd(\omega)$ by the definition of modality $\overline{\mathsf{K}}$ and the maximality of the set $hd(\omega)$. Thus, $hd(\omega) \vdash \mathsf{K}_D\mathsf{K}_D\neg\mathsf{B}_D\chi$ by Lemma 6.6. Then, $hd(\omega) \vdash \mathsf{K}_C\mathsf{K}_D\neg\mathsf{B}_D\chi$ by the Monotonicity axiom and because $D \subseteq C$. Thus, $\mathsf{K}_C\mathsf{K}_D\neg\mathsf{B}_D\chi \in hd(\omega)$ by the maximality of the set $hd(\omega)$. Hence, $\mathsf{K}_D\neg\mathsf{B}_D\chi \in X$ by the choice of set X. Thus, $\mathsf{K}_D\neg\mathsf{B}_D\chi \in X' = hd(\omega')$ by the choice of set X' and the choice of sequence ω'. Then, $\neg\mathsf{K}_D\neg\mathsf{B}_D\chi \notin hd(\omega')$ because set $hd(\omega')$ is consistent. Therefore, $\overline{\mathsf{K}}_D\mathsf{B}_D\chi \notin hd(\omega')$ by the definition of modality $\overline{\mathsf{K}}$, which contradicts the choice of formula $\overline{\mathsf{K}}_D\mathsf{B}_D\chi$.

Case II: $D \nsubseteq C$. Consider any $d_0 \in D \setminus C$. Thus, $\delta'(d_0) = \bot$ by equation (9). Also, $\delta'(d_0) = \chi$ because $d_0 \in D$. Thus, $\chi \equiv \bot$. Hence, formula $\neg\chi$ is a tautology. Therefore, $\neg\chi \in hd(\omega')$ by the maximality of set $hd(\omega')$. ⊠

This concludes the proof of the lemma. □

LEMMA 8.14. *For any outcome $\omega \in \Omega$, there is an initial state $\alpha \in I$ and a complete action profile $\delta \in \Delta^{\mathcal{A}}$ such that $(\alpha, \delta, \omega) \in P$.*

Proof. Let initial state α be the equivalence class of outcome ω with respect to the equivalence relation $\sim_{\mathcal{A}}$. Thus, $\omega \in \alpha$. Let δ be the complete action profile such that $\delta(a) = \perp$ for each $a \in \mathcal{A}$. To prove $(\alpha, \delta, \omega) \in P$, consider any formula $\overline{\mathsf{K}}_D\mathsf{B}_D\chi \in hd(\omega)$ such that $\delta(a) = \chi$ for each $a \in D$. By Definition 8.11, it suffices to show that $\neg\chi \in hd(\omega)$.

Case I: $D = \varnothing$. Thus, $\vdash \neg\mathsf{B}_D\chi$ by the None to Blame axiom. Hence, $\vdash \mathsf{K}_D\neg\mathsf{B}_D\chi$ by the Necessitation rule. Then, $\neg\mathsf{K}_D\neg\mathsf{B}_D\chi \notin hd(\omega)$ because set $hd(\omega)$ is consistent. Therefore, $\overline{\mathsf{K}}_D\mathsf{B}_D\chi \notin hd(\omega)$ by the definition of modality $\overline{\mathsf{K}}$, which contradicts the choice of formula $\overline{\mathsf{K}}_D\mathsf{B}_D\chi$.

Case II: $D \neq \varnothing$. Then, there is at least one agent $d_0 \in D$. Hence, $\chi = \delta(d_0) = \perp$ by the definition of the complete action profile δ. Then, $\neg\chi$ is a tautology. Thus, $\neg\chi \in hd(\omega)$ by the maximality of set $hd(\omega)$. □

Lemma 8.15. *For any $(\alpha, \delta, \omega) \in P$ and any $\neg\mathsf{K}_C\varphi \in hd(\omega)$, there is a play $(\alpha', \delta', \omega') \in P$ such that $\alpha \sim_C \alpha'$ and $\neg\varphi \in hd(\omega')$.*

Proof. Consider the set $X = \{\neg\varphi\} \cup \{\psi \mid \mathsf{K}_C\psi \in hd(\omega)\}$. First, we show that set X is consistent. Suppose the opposite. Then, there are formulae $\mathsf{K}_C\psi_1, \dots, \mathsf{K}_C\psi_n \in hd(\omega)$ such that $\psi_1, \dots, \psi_n \vdash \varphi$. Hence, $\mathsf{K}_C\psi_1, \dots, \mathsf{K}_C\psi_n \vdash \mathsf{K}_C\varphi$ by Lemma 6.5. Thus, $hd(\omega) \vdash \mathsf{K}_C\varphi$ because $\mathsf{K}_C\psi_1, \dots, \mathsf{K}_C\psi_n \in hd(\omega)$. Hence, $\neg\mathsf{K}_C\varphi \notin hd(\omega)$ because set $hd(\omega)$ is consistent, which contradicts the assumption of the lemma. Therefore, set X is consistent.

Let set X' be any maximal consistent extension of set X and ω' be the sequence $\omega :: C :: X'$. Note that $\omega' \in \Omega$ by Definition 8.1 and the choice of sets X and X'. Also, $\neg\varphi \in X \subseteq X' = hd(\omega')$ by the choice of sets X and X'.

By Lemma 8.14, there is an initial state $\alpha' \in I$ and a complete action profile δ' such that $(\alpha', \delta', \omega') \in P$. Note that $\omega \sim_C \omega'$ by Definition 8.2 and the choice of sequence ω'. Thus, $\alpha \sim_C \alpha'$ by Lemma 8.9. □

Lemma 8.16. *$(\alpha, \delta, \omega) \Vdash \varphi$ iff $\varphi \in hd(\omega)$ for each play $(\alpha, \delta, \omega) \in P$ and each formula $\varphi \in \Phi$.*

Proof. We prove the lemma by induction on complexity of formula φ. If φ is a propositional variable, then the lemma follows from Definition 4.3 and Definition 8.12. If formula φ is an implication or a negation then the required follow from the maximality and the consistency of set ω by Definition 4.3 in the standard way.

Assume that formula φ has the form $\mathsf{K}_C\psi$.

$(\Rightarrow)$: Let $\mathsf{K}_C\psi \notin hd(\omega)$. Thus, $\neg\mathsf{K}_C\psi \in hd(\omega)$ by the maximality of set $hd(\omega)$. Hence, by Lemma 8.15, there is a play $(\alpha', \delta', \omega') \in P$ such that $\alpha \sim_C \alpha'$ and $\neg\psi \in hd(\omega')$. Then, $\psi \notin hd(\omega')$ by the consistency of set $hd(\omega')$. Thus, $(\alpha', \delta', \omega') \nVdash \psi$ by the induction hypothesis. Therefore, $(\alpha, \delta, \omega) \nVdash \mathsf{K}_C\psi$ by Definition 4.3.

$(\Leftarrow)$: Let $\mathsf{K}_C\psi \in hd(\omega)$. Thus, $\psi \in hd(\omega')$ for any $\omega' \in \Omega$ such that $\omega \sim_C \omega'$, by Lemma 8.6. Hence, by the induction hypothesis, $(\alpha', \delta', \omega') \Vdash \psi$ for each play $(\alpha', \delta', \omega') \in P$ such that $\omega \sim_C \omega'$. Thus, $(\alpha', \delta', \omega') \Vdash \psi$ for each $(\alpha', \delta', \omega') \in P$ such that $\alpha \sim_C \alpha'$, by Lemma 8.9. Therefore, $(\alpha, \delta, \omega) \Vdash \mathsf{K}_C\psi$ by Definition 4.3.

Assume that formula φ has the form $\mathsf{B}_C\psi$.

$(\Rightarrow)$: Suppose $\mathsf{B}_C\psi \notin hd(\omega)$. First, consider the case when $\psi \notin hd(\omega)$. Then, $(\alpha, \delta, \omega) \nVdash \psi$ by the induction hypothesis. Thus, $(\alpha, \delta, \omega) \nVdash \mathsf{B}_C\psi$ by Definition 4.3.

Next, suppose $\psi \in hd(\omega)$. Observe that $\psi \to \mathsf{B}_C\psi \notin hd(\omega)$. Indeed, if $\psi \to \mathsf{B}_C\psi \in hd(\omega)$, then $hd(\omega) \vdash \mathsf{B}_C\psi$ by the Modus Ponens inference rule. Thus, $\mathsf{B}_C\psi \in hd(\omega)$ by the maximality of set $hd(\omega)$, which contradicts the assumption above.

Because $hd(\omega)$ is a maximal set, statement $\psi \to \mathsf{B}_C\psi \notin hd(\omega)$ implies that $\neg(\psi \to \mathsf{B}_C\psi) \in hd(\omega)$. Hence, by Lemma 8.13, for any action profile $s \in \Delta^C$, there is a play $(\alpha', \delta', \omega')$ such that $\alpha \sim_C \alpha'$ and $\psi \in hd(\omega')$. Thus, by the induction hypothesis, for any action profile $s \in \Delta^C$, there is a play $(\alpha', \delta', \omega')$ such that $\alpha \sim_C \alpha'$ and $(\alpha', \delta', \omega') \Vdash \psi$. Therefore, $(\alpha, \delta, \omega) \nVdash \mathsf{B}_C\psi$ by Definition 4.3.

$(\Leftarrow)$: Let $\mathsf{B}_C\psi \in hd(\omega)$. Hence, $hd(\omega) \vdash \psi$ by the Truth axiom. Thus, $\psi \in hd(\omega)$ by the maximality of the set $hd(\omega)$. Then, $(\alpha, \delta, \omega) \Vdash \psi$ by the induction hypothesis.

Next, let $s \in \Delta^C$ be the action profile of coalition C such that $s(a) = \psi$ for each agent $a \in C$. Consider any play $(\alpha', \delta', \omega') \in P$ such that $\alpha \sim_C \alpha'$ and $s =_C \delta'$. By Definition 4.3, it suffices to show that $(\alpha', \delta', \omega') \nVdash \psi$.

Indeed, by Lemma 6.3, assumption $\mathsf{B}_C\psi \in hd(\omega)$ implies that $hd(\omega) \vdash \overline{\mathsf{K}}_C\mathsf{B}_C\psi$. Thus, $hd(\omega) \vdash \mathsf{K}_C\overline{\mathsf{K}}_C\mathsf{B}_C\psi$ by the Negative introspection axiom, the Modus Ponens inference rule, and the definition of modality $\overline{\mathsf{K}}$. Hence, $\mathsf{K}_C\overline{\mathsf{K}}_C\mathsf{B}_C\psi \in hd(\omega)$ by the maximality of set $hd(\omega)$. Observe that $\omega \sim_C \omega'$ by Lemma 8.9 and the assumption $\alpha \sim_C \alpha'$. Thus, $\overline{\mathsf{K}}_C\mathsf{B}_C\psi \in hd(\omega')$ by Lemma 8.6.

Recall that $s(a) = \psi$ for each agent $a \in C$ by the choice of the action profile s. Also, $s =_C \delta'$ by the choice of the play $(\alpha', \delta', \omega')$. Hence, $\delta'(a) = \psi$ for each agent $a \in C$. Thus, $\neg\psi \in hd(\omega')$ by Definition 8.11 and because $\overline{\mathsf{K}}_C\mathsf{B}_C\psi \in hd(\omega')$. Then, $\psi \notin hd(\omega')$ the consistency of set $hd(\omega')$. Therefore, $(\alpha', \delta', \omega') \nVdash \psi$ by the induction hypothesis. □

Finally, we are now ready to state and to prove the strong completeness of our logical system.

Theorem 8.17. *If $X \nvdash \varphi$, then there is a game, and a play (α, δ, ω) of this game such that $(\alpha, \delta, \omega) \Vdash \chi$ for each $\chi \in X$ and $(\alpha, \delta, \omega) \nVdash \varphi$.*

Proof. Assume that $X \nvdash \varphi$. Hence, set $X \cup \{\neg\varphi\}$ is consistent. Let X_0 be any maximal consistent extension of set $X \cup \{\neg\varphi\}$ and let game $(I, \{\sim_a\}_{a \in \mathcal{A}}, \Delta, \Omega, P, \pi)$ be the canonical game $G(X_0)$. Also, let ω_0 be the single-element sequence X_0. Note that $\omega_0 \in \Omega$ by Definition 8.1. By Lemma 8.14, there is an initial state $\alpha \in I$ and a complete action profile $\delta \in \Delta^{\mathcal{A}}$ such that $(\alpha, \delta, \omega_0) \in P$. Hence, $(\alpha, \delta, \omega_0) \Vdash \chi$ for each $\chi \in X$ and $(\alpha, \delta, \omega_0) \Vdash \neg\varphi$ by Lemma 8.16 and the choice of set X_0. Therefore, $(\alpha, \delta, \omega_0) \nVdash \varphi$ by Definition 4.3. □

9 CONCLUSION

In this paper we proposed a definition of blameworthiness in strategic games with imperfect information and gave a sound and complete logical system that captures the interplay between distributed knowledge and blameworthiness modalities. In [38], we proposed the notion of a second-order known-how. A coalition is said to have second-order know-how knowledge if the coalition knows how *another coalition* can achieve the goal. In the future work, we plan to explore the notion of second-order blameworthiness in strategic games, which refers to the situation when one coalition knew how another could have prevented the outcome.

REFERENCES

[1] Thomas Ågotnes and Natasha Alechina. 2016. Coalition Logic with Individual, Distributed and Common Knowledge. *Journal of Logic and Computation* (2016). https://doi.org/10.1093/logcom/exv085 exv085.

[2] Thomas Ågotnes, Philippe Balbiani, Hans van Ditmarsch, and Pablo Seban. 2010. Group announcement logic. *Journal of Applied Logic* 8, 1 (2010), 62 – 81. https://doi.org/10.1016/j.jal.2008.12.002

[3] Thomas Ågotnes, Wiebe van der Hoek, and Michael Wooldridge. 2009. Reasoning about coalitional games. *Artificial Intelligence* 173, 1 (2009), 45 – 79. https://doi.org/10.1016/j.artint.2008.08.004

[4] Natasha Alechina, Joseph Y Halpern, and Brian Logan. 2017. Causality, responsibility and blame in team plans. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1091–1099.

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

[6] Vitaliy Batusov and Mikhail Soutchanski. 2018. Situation Calculus Semantics for Actual Causality. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*.

[7] Francesco Belardinelli. 2014. Reasoning about Knowledge and Strategies: Epistemic Strategy Logic. In *Proceedings 2nd International Workshop on Strategic Reasoning, SR 2014, Grenoble, France, April 5-6, 2014 (EPTCS)*, Vol. 146. 27–33.

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

[9] Stefano Borgo. 2007. Coalitions in Action Logic. In *20th International Joint Conference on Artificial Intelligence*. 1822–1827.

[10] Jan Broersen. 2008. A Logical Analysis of the Interaction between 'Obligation-to-do' and 'Knowingly Doing'. In *International Conference on Deontic Logic in Computer Science*. Springer, 140–154.

[11] Jan Broersen, Andreas Herzig, and Nicolas Troquard. 2009. What groups do, can do, and know they can do: an analysis in normal modal logics. *Journal of Applied Non-Classical Logics* 19, 3 (2009), 261–289. https://doi.org/10.3166/jancl.19.261-289

[12] Fiery Cushman. 2015. Deconstructing intent to reconstruct morality. *Current Opinion in Psychology* 6 (2015), 97–103.

[13] Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. 1995. *Reasoning about knowledge*. MIT Press, Cambridge, MA. xiv+477 pages.

[14] Raul Fervari, Andreas Herzig, Yanjun Li, and Yanjing Wang. 2017. Strategically knowing how. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 1031–1038.

[15] Lloyd Fields. 1994. Moral Beliefs and Blameworthiness: Introduction. *Philosophy* 69, 270 (1994), 397–415.

[16] John Martin Fischer and Mark Ravizza. 2000. *Responsibility and control: A theory of moral responsibility*. Cambridge University Press.

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

[18] Rustam Galimullin and Natasha Alechina. 2017. Coalition and Group Announcement Logic. In *Proceedings Sixteenth Conference on Theoretical Aspects of Rationality and Knowledge (TARK) 2017, Liverpool, UK, 24-26 July 2017*. 207–220.

[19] Michelle Gant. 2018. Millennials being blamed for decline of American cheese. *Fox News* (11 October 11, 2018). www.foxnews.com/food-drink/millennials-kraft-american-cheese-sales-decline.amp.

[20] Valentin Goranko. 2001. Coalition games and alternating temporal logics. In *Proceedings of the 8th conference on Theoretical aspects of rationality and knowledge*. Morgan Kaufmann Publishers Inc., 259–272.

[21] Valentin Goranko and Sebastian Enqvist. 2018. Socially Friendly and Group Protecting Coalition Logics. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 372–380.

[22] Valentin Goranko, Wojciech Jamroga, and Paolo Turrini. 2013. Strategic games and truly playable effectivity functions. *Autonomous Agents and Multi-Agent Systems* 26, 2 (01 Mar 2013), 288–314. https://doi.org/10.1007/s10458-012-9192-y

[23] Joseph Y Halpern. 2016. *Actual causality*. MIT Press.

[24] Joseph Y Halpern. 2017. *Reasoning about uncertainty*. MIT press.

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

[26] John Horty and Eric Pacuit. 2017. Action types in stit semantics. *The Review of Symbolic Logic* (2017), 1–21.

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

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

[29] American Law Institute. 1985 Print. *Model Penal Code: Official Draft and Explanatory Notes. Complete Text of Model Penal Code as Adopted at the 1962 Annual Meeting of the American Law Institute at Washington, D.C., May 24, 1962.* The Institute.

[30] Wojciech Jamroga and Thomas Ågotnes. 2007. Constructive knowledge: what agents can achieve under imperfect information. *Journal of Applied Non-Classical Logics* 17, 4 (2007), 423–475. https://doi.org/10.3166/jancl.17.423-475

[31] Wojciech Jamroga and Wiebe van der Hoek. 2004. Agents that know how to play. *Fundamenta Informaticae* 63, 2-3 (2004), 185–219.

[32] Emiliano Lorini and François Schwarzentruber. 2011. A logic for reasoning about counterfactual emotions. *Artificial Intelligence* 175, 3 (2011), 814.

[33] Elinor Mason. 2015. Moral ignorance and blameworthiness. *Philosophical Studies* 172, 11 (2015), 3037–3057.

[34] Pavel Naumov and Kevin Ros. 2018. Strategic Coalitions in Systems with Catastrophic Failures (extended abstract). In *Proceedings of the 16th International Conference on Principles of Knowledge Representation and Reasoning*.

[35] Pavel Naumov and Jia Tao. 2017. Coalition Power in Epistemic Transition Systems. In *Proceedings of the 2017 International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 723–731.

[36] Pavel Naumov and Jia Tao. 2017. Together We Know How to Achieve: An Epistemic Logic of Know-How. In *16th conference on Theoretical Aspects of Rationality and Knowledge (TARK), July 24-26, 2017, EPTCS 251*. 441–453.

[37] Pavel Naumov and Jia Tao. 2018. Blameworthiness in Strategic Games. *arXiv:1809.05485* (2018).

[38] Pavel Naumov and Jia Tao. 2018. Second-Order Know-How Strategies. In *Proceedings of the 2018 International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 390–398.

[39] Pavel Naumov and Jia Tao. 2018. Strategic Coalitions with Perfect Recall. In *Proceedings of Thirty-Second AAAI Conference on Artificial Intelligence*.

[40] Pavel Naumov and Jia Tao. 2018. Together We Know How to Achieve: An Epistemic Logic of Know-How. *Artificial Intelligence* 262 (2018), 279 – 300. https://doi.org/10.1016/j.artint.2018.06.007

[41] Shaun Nichols and Joshua Knobe. 2007. Moral responsibility and determinism: The cognitive science of folk intuitions. *Nous* 41, 4 (2007), 663–685.

[42] Grigory K Olkhovikov and Heinrich Wansing. 2018. Inference as doxastic agency. Part I: The basics of justification stit logic. *Studia Logica* (2018), 1–28.

[43] Marc Pauly. 2001. *Logic for Social Software*. Ph.D. Dissertation. Institute for Logic, Language, and Computation.

[44] Marc Pauly. 2002. A Modal Logic for Coalitional Power in Games. *Journal of Logic and Computation* 12, 1 (2002), 149–166. https://doi.org/10.1093/logcom/12.1.149

[45] Luigi Sauro, Jelle Gerbrandy, Wiebe van der Hoek, and Michael Wooldridge. 2006. Reasoning About Action and Cooperation. In *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS '06)*. ACM, New York, NY, USA, 185–192. https://doi.org/10.1145/1160633.1160663

[46] Peter Singer and Maya Eddon. 2013. Moral responsibility, problem of. *Encyclopædia Britannica* (2013). https://www.britannica.com/topic/problem-of-moral-responsibility.

[47] Johan van Benthem. 2001. Games in Dynamic-Epistemic Logic. *Bulletin of Economic Research* 53, 4 (2001), 219–248. https://doi.org/10.1111/1467-8586.00133

[48] Wiebe van der Hoek and Michael Wooldridge. 2005. On the logic of cooperation and propositional control. *Artificial Intelligence* 164, 1 (2005), 81 – 119.

[49] Yanjing Wang. 2015. A logic of knowing how. In *Logic, Rationality, and Interaction*. Springer, 392–405.

[50] Yanjing Wang. 2016. A logic of goal-directed knowing how. *Synthese* (2016), 1–21.

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

[52] Ming Xu. 1998. Axioms for deliberative STIT. *Journal of Philosophical Logic* 27, 5 (1998), 505–552.