# Duty to Warn in Strategic Games

Pavel Naumov
Tulane University
New Orleans, Louisiana
pgn2@cornell.edu

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

## ABSTRACT

The paper investigates the second-order blameworthiness or duty to warn modality "one coalition knew how another coalition could have prevented an outcome". The main technical result is a sound and complete logical system that describes the interplay between the distributed knowledge and the duty to warn modalities.

## KEYWORDS

epistemic logic, blameworthiness, responsibility, completeness

## 1 INTRODUCTION

On October 27, 1969, Prosenjit Poddar, an Indian graduate student from the University of California, Berkeley, came to the parents' house of Tatiana Tarasoff, an undergraduate student who recently immigrated from Russia. After a brief conversation, he pulled out a gun and unloaded it into her torso, then stabbed her eight times with a 13-inch butcher knife, walked into the house and called the police. Tarasoff was pronounced dead on arrival at the hospital [2].

In this paper we study the notion of blameworthiness. This notion is usually defined through the principle of alternative possibilities: an agent (or a coalition of agents) is blamable for $\varphi$ if $\varphi$ is true and the agent had a strategy to prevent it [8, 19]. This definition is also referred to as the counterfactual definition of blameworthiness [5]. In our case, Poddar is blamable for the death of Tatiana because he could have taken actions (to refrain from shooting and stabbing her) that would have prevented her death. He was found guilty of second-degree murder and sentenced to five years [2]. The principle of alternative possibilities, sometimes referred to as "counterfactual possibility" [5], is also used to define causality [3, 9, 10]. A sound and complete axiomatization of modality "statement $\varphi$ is true and coalition $C$ had a strategy to prevent $\varphi$" is proposed in [15]. In related works, Xu [20] and Broersen, Herzig, and Troquard [4] axiomatized modality "took actions that unavoidably resulted in $\varphi$" in the cases of single agents and coalitions respectively.

According to the principle of alternative possibilities, Poddar is not the only one who is blamable for Tatiana's death. Indeed, Tatiana's parents could have asked for a temporary police protection, hired a private bodyguard, or taken Tatiana on a long vacation outside of California. Each of these actions is likely to prevent Tatiana's death. Thus, by applying the principle of alternative possibilities directly, we have to conclude that her parents should be blamed for Tatiana's death. However, the police is unlikely to provide lifetime protection; the parents' resources can only be used to hire a bodyguard for a limited period time; and any vacation will have to end. These measures would only work if they knew an approximate time of a likely attack on their daughter. Without this crucial information, they had a strategy to prevent her death, but they did not know what this strategy was. If an agent has a strategy to achieve a certain outcome, knows that it has a strategy, and knows what this strategy is, then we say that the agent has a *know-how strategy*. Axiomatic systems for know-how strategies have been studied before [1, 7, 11, 13, 14, 16]. In a setting with imperfect information, it is natural to modify the principle of alternative possibilities to require an agent or a coalition to have a know-how strategy to prevent. In our case, parents had many different strategies that included taking vacations in different months. They did not know that a vacation in October would have prevented Tatiana's death. Thus, they cannot be blamed for her death according to the modified version of the principle of alternative possibilities. We write this as $\neg\mathsf{B}_{\text{parents}}$("Tatiana is killed").

Although Tatiana's parents did not know how to prevent her death, Dr. Lawrence Moore did. He was a psychiatrist who treated Poddar at the University of California mental clinic. Poddar told Moore how he met Tatiana at the University international student house, how they started to date and how depressed Poddar became when Tatiana lost romantic interest in him. Less than two months before the tragedy, Poddar shared with the doctor his intention to buy a gun and to murder Tatiana. Dr. Moore reported this information to the University campus police. Since the University knew that Poddar was at the peak of his depression, they could estimate the possible timing of the attack. Thus, the University knew what actions the parents could take to prevent the tragedy. In general, if a coalition $C$ knows how a coalition $D$ can achieve a certain outcome, then coalition $D$ has a *second-order know-how* strategy to achieve the outcome. This class of strategies and a complete logical system that describes its properties were proposed in [12]. We write $\mathsf{B}^D_C\varphi$ if *$\varphi$ is true and coalition $C$ knew how coalition $D$ could have prevented $\varphi$*. In our case, $\mathsf{B}^{\text{parents}}_{\text{university}}$("Tatiana is killed").

After Tatiana's death, her parents sued the University. In 1976 the California Supreme Court ruled that "When a therapist determines, or pursuant to the standards of his profession should determine, that his patient presents a serious danger of violence to another, he incurs an obligation to use reasonable care to protect the intended victim against such danger. The discharge of this duty may require the therapist to take one or more of various steps, depending upon the nature of the case. Thus it may call for him to warn the intended victim or others likely to apprise the victim of the danger, to notify the police, or to take whatever other steps are reasonably necessary under the circumstances." [18]. In other words, the California Supreme Court ruled that in this case the duty to warn is not only a moral obligation but a legal one as well. In

*Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 2020, Auckland, New Zealand*

https://doi.org/doi

this paper we propose a sound and complete logical system that describes the interplay between the distributed knowledge modality $\mathsf{K}_C$ and the second-order blameworthiness or *duty to warn* modality $\mathsf{B}^D_C$. The (first-order) blameworthiness modality $\mathsf{B}_C\varphi$ mentioned earlier could be viewed as an abbreviation for $\mathsf{B}^C_C\varphi$. For example, $\mathsf{B}_{\text{Poddar}}$("Tatiana is killed") because Poddar knew how he himself could prevent Tatiana's death.

The paper is organized as follows. In the next section we introduce and discuss the formal syntax and semantics of our logical system. In Section 3 we list axioms and compare them to those in the related logical systems. Section 4 gives examples of formal proofs in our system. Section 5 and Section 6 contain the proofs of the soundness and the completeness, respectively. Section 7 concludes.

## 2 SYNTAX AND SEMANTICS

In this section we introduce the formal syntax and semantics of our logical system. We assume a fixed set of propositional variables and a fixed set of agents $\mathcal{A}$. By a coalition we mean any subset of $\mathcal{A}$. The language $\Phi$ of our logical system is defined by grammar:

$$\varphi := p \mid \neg\varphi \mid \varphi \to \varphi \mid \mathsf{K}_C\varphi \mid \mathsf{B}^D_C\varphi,$$

where $C$ and $D$ are arbitrary coalitions. Boolean connectives $\bot$, $\wedge$, and $\vee$ are defined through $\neg$ and $\to$ in the usual way. By $\overline{\mathsf{K}}_C\varphi$ we denote the formula $\neg\mathsf{K}_C\neg\varphi$ and by $X^Y$ the set of all functions from set $Y$ to set $X$.

*Definition 2.1.* 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 the set of initial states $I$, for each agent $a \in \mathcal{A}$,
(3) $\Delta$ is a set of "actions",
(4) $\Omega$ is a set of "outcomes",
(5) a set of "plays" $P$ is an arbitrary set of tuples $(\alpha, \delta, \omega)$ such that $\alpha \in I$, $\delta \in \Delta^{\mathcal{A}}$, and $\omega \in \Omega$. Furthermore, we assume that for each initial state $\alpha \in I$ and each function $\delta \in \Delta^{\mathcal{A}}$, there is at least one outcome $\omega \in \Omega$ such that $(\alpha, \delta, \omega) \in P$,
(6) $\pi(p) \subseteq P$ for each propositional variable $p$.

By a complete (action) profile we mean any function $\delta \in \Delta^{\mathcal{A}}$ that maps agents in $\mathcal{A}$ into actions in $\Delta$. By an (action) profile of a coalition $C$ we mean any function from set $\Delta^C$.

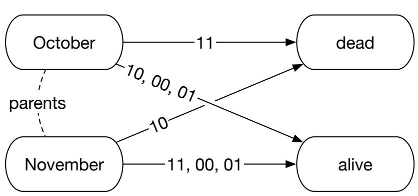


**Figure 1: Poddar's actions: not attack (0) or attack (1). Parents' actions: take vacation in October (0) or November (1).**

Figure 1 depicts a diagram of the game for the Tarasoff case. It shows two possible initial states: October and November that represent two possible months with the peak of Poddar's depression. The actual initial state was October, which was known to the University, but not to Tatiana's parents. In other words, the University could distinguish these two states, but the parents could not. We show the indistinguishability relation by dashed lines. At the peak of his depression, agent Poddar might decide not to attack Tatiana (action 0) or to attack her (action 1). Parents, whom we represent by a single agent for the sake of simplicity, might decide to take vacation in October (action 0) or November (action 1). Thus, in our example, $\Delta = \{0, 1\}$. Set $\Omega$ consists of outcomes *dead* and *alive*. Recall that a complete action profile is a function from agents into actions. Since in our case there are only two agents (Poddar and parents), we write action profiles as $xy$ where $x \in \{0, 1\}$ is an action of Poddar and $y \in \{0, 1\}$ is an action of the parents. The plays of the game are all possible valid combinations of an initial state, a complete action profile, and an outcome. The plays are represented in the diagram by directed edges. For example, the directed edge from initial state October to outcome *dead* is labeled with action profile 11. This means that $(\text{October}, 11, \textit{dead}) \in P$. In other words, if the peak of depression is in October, Poddar decides to attack (1), and the parents take vacation in November (1), then Tatiana is dead. Multiple labels on the same edge of the diagram represent multiple plays with the same initial state and the same outcome.

Function $\pi$ specifies the meaning of propositional variables. Namely, $\pi(p)$ is the set of all plays for which proposition $p$ is true.

Next is the core definition of this paper. Its item 5 formally defines the semantics of modality $\mathsf{B}^D_C$. Traditionally, in modal logic the satisfiability $\Vdash$ is defined as a relation between a state and a formula. This approach is problematic in the case of the blameworthiness modality because this modality refers to two different states: $\mathsf{B}^D_C\varphi$ if statement $\varphi$ is true in *the current* state and coalition $C$ knew how coalition $D$ could have prevented $\varphi$ in *the previous* state. In other words, the meaning of formula $\mathsf{B}^D_C\varphi$ depends not only on the current state, but on the previous one as well. We resolve this issue by defining the satisfiability as a relation between a *play* and a formula, where a play is a triple consisting of the previous state $\alpha$, the complete action profile $\delta$, and an outcome (state) $\omega$. We distinguish initial states from outcomes to make the presentation more elegant. Otherwise, this distinction is not significant.

We write $\omega \sim_C \omega'$ if $\omega \sim_a \omega'$ for each agent $a \in C$. We also write $f =_X g$ if $f(x) = g(x)$ for each element $x \in X$.

*Definition 2.2.* 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 $(\alpha, \delta, \omega) \in \pi(p)$,
(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 $(\alpha', \delta', \omega') \in P$ such that $\alpha \sim_C \alpha'$,
(5) $(\alpha, \delta, \omega) \Vdash \mathsf{B}^D_C\varphi$ if $(\alpha, \delta, \omega) \Vdash \varphi$ and there is a profile $s \in \Delta^D$ such that for each play $(\alpha', \delta', \omega') \in P$, if $\alpha \sim_C \alpha'$ and $s =_D \delta'$, then $(\alpha', \delta', \omega') \nVdash \varphi$.

Going back to our running example,

$$(\text{October}, 11, \textit{dead}) \Vdash \mathsf{B}^{\text{parents}}_{\text{university}}(\text{"Tatiana is killed"})$$

because $(\text{October}, 11, \textit{dead}) \Vdash$ "Tatiana is killed" and

$$(\alpha', \delta', \omega') \nVdash (\text{"Tatiana is killed"})$$

for each play $(\alpha', \delta', \omega') \in P$ such that $\alpha' \sim_{\text{university}}$ October and $\delta'(\text{parents}) = 0$.

Because the satisfiability is defined as a relation between plays and formulae, one can potentially talk about two forms of knowledge *about a play* in our system: *a priori* knowledge in the initial state and *a posteriori* knowledge in the outcome. The knowledge captured by the modality K as well as the knowledge implicitly referred to by the modality B, see item (5) of Definition 2.2, is *a priori* knowledge about a play. In order to define a posteriori knowledge in our setting, one would need to add an indistinguishability relation on outcomes to Definition 2.1. We do not consider a posteriori knowledge because one should not be blamed for something that the person only knows how to prevent *post-factum*.

Since we define the second-order blameworthiness using distributed knowledge, if a coalition $C$ is blamable for not warning coalition $D$, then any superset $C' \supseteq C$ could be blamed for not warning $D$. One might argue that the definition of blameworthiness modality $\mathsf{B}_C^D$ should include a minimality condition on the coalition $C$. We do not include this condition in item (5) of Definition 2.2, because there are several different ways to phrase the minimality, all of which could be expressed through our basic modality $\mathsf{B}_C^D$.

First of all, we can say that $C$ is the minimal coalition among those coalitions that knew how $D$ could have prevented $\varphi$. Let us denote this modality by $[1]_C^D\varphi$. It can be expressed through $\mathsf{B}_C^D$ as:

$$[1]_C^D\varphi \equiv \mathsf{B}_C^D\varphi \wedge \neg \bigvee_{E \subsetneq C} \mathsf{B}_E^D\varphi.$$

Second, we can say that $C$ is the minimal coalition that knew how *somebody* could have prevented $\varphi$:

$$[2]_C^D\varphi \equiv \mathsf{B}_C^D\varphi \wedge \neg \bigvee_{E \subsetneq C} \bigvee_{F \subseteq \mathcal{A}} \mathsf{B}_E^F\varphi.$$

Third, we can say that $C$ is the minimal coalition that knew how *the smallest* coalition $D$ could have prevented $\varphi$:

$$[3]_C^D\varphi \equiv \mathsf{B}_C^D\varphi \wedge \neg \bigvee_{E \subseteq \mathcal{A}} \bigvee_{F \subsetneq D} \mathsf{B}_E^F\varphi \wedge \neg \bigvee_{E \subsetneq C} \mathsf{B}_E^D\varphi.$$

Finally, we can say that $C$ is the minimal coalition that knew how *some smallest* coalition could have prevented $\varphi$:

$$[4]_C\varphi \equiv \bigvee_{D \subseteq \mathcal{A}} \left( \mathsf{B}_C^D\varphi \wedge \neg \bigvee_{E \subseteq \mathcal{A}} \bigvee_{F \subsetneq D} \mathsf{B}_E^F\varphi \wedge \neg \bigvee_{E \subsetneq C} \mathsf{B}_E^D\varphi \right).$$

The choice of the minimality condition depends on the specific situation. Instead of making a choice between several possible alternatives, in this paper we study the basic blameworthiness modality without a minimality condition through which modalities $[1]_C^D\varphi$, $[2]_C^D\varphi$, $[3]_C^D\varphi$, $[4]_C\varphi$, and possibly others could be defined.

## 3 AXIOMS

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

(1) Truth: $\mathsf{K}_C\varphi \to \varphi$ and $\mathsf{B}_C^D\varphi \to \varphi$,
(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: $\mathsf{K}_C\varphi \to \mathsf{K}_E\varphi$ and $\mathsf{B}_C^D\varphi \to \mathsf{B}_E^F\varphi$, where $C \subseteq E$ and $D \subseteq F$,
(5) None to Act: $\neg\mathsf{B}_C^{\varnothing}\varphi$,
(6) Joint Responsibility: if $D \cap F = \varnothing$, then $\overline{\mathsf{K}}_C\mathsf{B}_C^D\varphi \wedge \overline{\mathsf{K}}_E\mathsf{B}_E^F\psi \to (\varphi \vee \psi \to \mathsf{B}_{C\cup E}^{D\cup F}(\varphi \vee \psi))$,
(7) Strict Conditional: $\mathsf{K}_C(\varphi \to \psi) \to (\mathsf{B}_C^D\psi \to (\varphi \to \mathsf{B}_C^D\varphi))$,
(8) Introspection of Blameworthiness: $\mathsf{B}_C^D\varphi \to \mathsf{K}_C(\varphi \to \mathsf{B}_C^D\varphi)$.

The Truth, the Distributivity, the Negative Introspection, and the Monotonicity axioms for modality K are the standard axioms from the epistemic logic S5 for distributed knowledge [6]. The Truth axiom for modality B states that a coalition can only be blamed for something that has actually happened. The Monotonicity axiom for modality B captures the fact that both distributed knowledge and coalition power are monotonic.

The None to Act axiom is true because the empty coalition has only one action profile. Thus, if the empty coalition can prevent $\varphi$, then $\varphi$ would have to be false on the current play. This axiom is similar to the None to Blame axiom $\neg\mathsf{B}_{\varnothing}\varphi$ in [15].

The Joint Responsibility axiom shows how the blame of two separate coalitions can be combined into the blame of their union. This axiom is closely related to Marc Pauly [17] Cooperation axiom, which is also used in coalitional modal logics of know-how [1, 11, 13, 14] and second-order know-how [12]. We formally prove the soundness of this axiom in Lemma 5.1.

Strict conditional $\mathsf{K}_C(\varphi \to \psi)$ states that formula $\varphi$ is known to $C$ to imply $\psi$. By contraposition, coalition $C$ knows that if $\psi$ is prevented, then $\varphi$ is also prevented. The Strict Conditional axiom states that if $C$ could be second-order blamed for $\psi$, then it should also be second-order blamed for $\varphi$ as long as $\varphi$ is true. A similar axiom is present in [15].

Finally, the Introspection of Blameworthiness axiom says that if coalition $C$ is second-order blamed for $\varphi$, then $C$ knows that it is second-order blamed for $\varphi$ as long as $\varphi$ is true. A similar Strategic Introspection axiom for second-order know-how modality is present in [12].

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

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

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

Lemma 3.1. *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 next lemma capture a well-known property of S5 modality. Its proof could be found, for example, in [12].

Lemma 3.2 (Positive Introspection). $\vdash \mathsf{K}_C\varphi \to \mathsf{K}_C\mathsf{K}_C\varphi$.

## 4 EXAMPLES OF DERIVATIONS

The soundness of our logical system is established in the next section. Here we prove several lemmas about our formal system that will be used later in the proof of the completeness.

Lemma 4.1. $\vdash \overline{\mathsf{K}}_C\mathsf{B}^D_C\varphi \to (\varphi \to \mathsf{B}^D_C\varphi)$.

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

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

At the same time, by the Negative Introspection axiom:

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

Then, by the laws of propositional reasoning,

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

Thus, by the law of contrapositive,

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

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

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

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

Lemma 4.2. *If* $\vdash \varphi \leftrightarrow \psi$, *then* $\vdash \mathsf{B}^D_C\varphi \to \mathsf{B}^D_C\psi$.

Proof. By the Strict Conditional axiom,

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

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, $\vdash \mathsf{B}^D_C\varphi \to (\psi \to \mathsf{B}^D_C\psi)$. Then, by the laws of propositional reasoning,

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

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

Lemma 4.3. $\varphi \vdash \overline{\mathsf{K}}_C\varphi$.

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 4.4. *For any integer* $n \geq 0$,

$$\{\overline{\mathsf{K}}_{E_i}\mathsf{B}^{F_i}_{E_i}\chi_i\}_{i=1}^n, \chi_1 \vee \dots \vee \chi_n \vdash \mathsf{B}^{F_1\cup\dots\cup F_n}_{E_1\cup\dots\cup E_n}(\chi_1 \vee \dots \vee \chi_n),$$

*where sets* $F_1, \dots, F_n$ *are pairwise disjoint.*

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 $\perp$. Hence, the statement of the lemma, $\perp \vdash \mathsf{B}^{\varnothing}_{\varnothing}\perp$, is provable in the propositional logic.

Next, assume that $n = 1$. Then, from Lemma 4.1 using Modus Ponens rule twice, we get $\overline{\mathsf{K}}_{E_1}\mathsf{B}^{F_1}_{E_1}\chi_1, \chi_1 \vdash \mathsf{B}^{F_1}_{E_1}\chi_1$.

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

$$\overline{\mathsf{K}}_{E_1\cup\dots\cup E_{n-1}}\mathsf{B}^{F_1\cup\dots\cup F_{n-1}}_{E_1\cup\dots\cup E_{n-1}}(\chi_1 \vee \dots \vee \chi_{n-1}), \overline{\mathsf{K}}_{E_n}\mathsf{B}^{F_n}_{E_n}\chi_n,$$
$$\chi_1 \vee \dots \vee \chi_{n-1} \vee \chi_n \vdash \mathsf{B}^{F_1\cup\dots\cup F_{n-1}\cup F_n}_{E_1\cup\dots\cup E_{n-1}\cup E_n}(\chi_1 \vee \dots \vee \chi_{n-1} \vee \chi_n).$$

Hence, by Lemma 4.3,

$$\mathsf{B}^{F_1\cup\dots\cup F_{n-1}}_{E_1\cup\dots\cup E_{n-1}}(\chi_1 \vee \dots \vee \chi_{n-1}), \overline{\mathsf{K}}_{E_n}\mathsf{B}^{F_n}_{E_n}\chi_n, \chi_1 \vee \dots \vee \chi_{n-1} \vee \chi_n$$
$$\vdash \mathsf{B}^{F_1\cup\dots\cup F_{n-1}\cup F_n}_{E_1\cup\dots\cup E_{n-1}\cup E_n}(\chi_1 \vee \dots \vee \chi_{n-1} \vee \chi_n).$$

At the same time, by the induction hypothesis,

$$\{\overline{\mathsf{K}}_{E_i}\mathsf{B}^{F_i}_{E_i}\chi_i\}_{i=1}^{n-1}, \chi_1 \vee \dots \vee \chi_{n-1} \vdash \mathsf{B}^{F_1\cup\dots\cup F_{n-1}}_{E_1\cup\dots\cup E_{n-1}}(\chi_1 \vee \dots \vee \chi_{n-1}).$$

Thus,

$$\{\overline{\mathsf{K}}_{E_i}\mathsf{B}^{F_i}_{E_i}\chi_i\}_{i=1}^n, \chi_1 \vee \dots \vee \chi_{n-1}, \chi_1 \vee \dots \vee \chi_{n-1} \vee \chi_n$$
$$\vdash \mathsf{B}^{F_1\cup\dots\cup F_{n-1}\cup F_n}_{E_1\cup\dots\cup E_{n-1}\cup E_n}(\chi_1 \vee \dots \vee \chi_{n-1} \vee \chi_n).$$

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,

$$\{\overline{\mathsf{K}}_{E_i}\mathsf{B}^{F_i}_{E_i}\chi_i\}_{i=1}^n, \chi_1 \vee \dots \vee \chi_{n-1}$$
$$\vdash \mathsf{B}^{F_1\cup\dots\cup F_{n-1}\cup F_n}_{E_1\cup\dots\cup E_{n-1}\cup E_n}(\chi_1 \vee \dots \vee \chi_{n-1} \vee \chi_n). \tag{2}$$

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

$$\overline{\mathsf{K}}_{E_1}\mathsf{B}^{F_1}_{E_1}\chi_1, \overline{\mathsf{K}}_{E_2\cup\dots\cup E_n}\mathsf{B}^{F_2\cup\dots\cup F_n}_{E_2\cup\dots\cup E_n}(\chi_2 \vee \dots \vee \chi_n),$$
$$\chi_1 \vee (\chi_2 \vee \dots \vee \chi_n) \vdash \mathsf{B}^{F_1\cup\dots\cup F_{n-1}\cup F_n}_{E_1\cup\dots\cup E_{n-1}\cup E_n}(\chi_1 \vee (\chi_2 \vee \dots \vee \chi_n)).$$

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 4.2,

$$\overline{\mathsf{K}}_{E_1}\mathsf{B}^{F_1}_{E_1}\chi_1, \overline{\mathsf{K}}_{E_2\cup\dots\cup E_n}\mathsf{B}^{F_2\cup\dots\cup F_n}_{E_2\cup\dots\cup E_n}(\chi_2 \vee \dots \vee \chi_n),$$
$$\chi_1 \vee \chi_2 \vee \dots \vee \chi_n \vdash \mathsf{B}^{F_1\cup\dots\cup F_{n-1}\cup F_n}_{E_1\cup\dots\cup E_{n-1}\cup E_n}(\chi_1 \vee \chi_2 \vee \dots \vee \chi_n).$$

Hence, by Lemma 4.3,

$$\overline{\mathsf{K}}_{E_1}\mathsf{B}^{F_1}_{E_1}\chi_1, \mathsf{B}^{F_2\cup\dots\cup F_n}_{E_2\cup\dots\cup E_n}(\chi_2 \vee \dots \vee \chi_n), \chi_1 \vee \chi_2 \vee \dots \vee \chi_n$$
$$\vdash \mathsf{B}^{F_1\cup\dots\cup F_{n-1}\cup F_n}_{E_1\cup\dots\cup E_{n-1}\cup E_n}(\chi_1 \vee \chi_2 \vee \dots \vee \chi_n).$$

At the same time, by the induction hypothesis,

$$\{\overline{\mathsf{K}}_{E_i}\mathsf{B}^{F_i}_{E_i}\chi_i\}_{i=2}^n, \chi_2 \vee \dots \vee \chi_n \vdash \mathsf{B}^{F_2\cup\dots\cup F_n}_{E_2\cup\dots\cup E_n}(\chi_2 \vee \dots \vee \chi_n).$$

Thus,

$$\{\overline{\mathsf{K}}_{E_i}\mathsf{B}^{F_i}_{E_i}\chi_i\}_{i=1}^n, \chi_2 \vee \dots \vee \chi_n, \chi_1 \vee \chi_2 \vee \dots \vee \chi_n$$
$$\vdash \mathsf{B}^{F_1\cup\dots\cup F_{n-1}\cup F_n}_{E_1\cup\dots\cup E_{n-1}\cup E_n}(\chi_1 \vee \chi_2 \vee \dots \vee \chi_n).$$

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,

$$\{\overline{\mathsf{K}}_{E_i}\mathsf{B}^{F_i}_{E_i}\chi_i\}_{i=1}^n, \chi_2 \vee \dots \vee \chi_n$$
$$\vdash \mathsf{B}^{F_1\cup\dots\cup F_{n-1}\cup F_n}_{E_1\cup\dots\cup E_{n-1}\cup E_n}(\chi_1 \vee \chi_2 \vee \dots \vee \chi_n). \tag{3}$$

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

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

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

$$\{\overline{\mathsf{K}}_{E_i}\mathsf{B}^{F_i}_{E_i}\chi_i\}_{i=1}^n, \chi_1 \vee \cdots \vee \chi_n \vdash \mathsf{B}^{F_1\cup\cdots\cup F_n}_{E_1\cup\cdots\cup E_n}(\chi_1 \vee \cdots \vee \chi_n).$$

by the laws of propositional reasoning. □

Our last example rephrases Lemma 4.4 into the form which is used in the proof of the completeness.

Lemma 4.5. *For any $n \geq 0$, any sets $E_1, \dots, E_n \subseteq C$, and any pairwise disjoint sets $F_1, \dots, F_n \subseteq D$,*

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

Proof. Let $X = \{\overline{\mathsf{K}}_{E_i}\mathsf{B}^{F_i}_{E_i}\chi_i\}_{i=1}^n$. Then, by Lemma 4.4,

$$X, \chi_1 \vee \cdots \vee \chi_n \vdash \mathsf{B}^{F_1\cup\cdots\cup F_n}_{E_1\cup\cdots\cup E_n}(\chi_1 \vee \cdots \vee \chi_n).$$

Hence, by the Monotonicity axiom,

$$X, \chi_1 \vee \cdots \vee \chi_n \vdash \mathsf{B}^D_C(\chi_1 \vee \cdots \vee \chi_n).$$

Thus, $\quad X, \varphi, \varphi \to \chi_1 \vee \cdots \vee \chi_n \vdash \mathsf{B}^D_C(\chi_1 \vee \cdots \vee \chi_n)$

by the Modus Ponens inference rule. Hence, by the Truth axiom,

$$X, \varphi, \mathsf{K}_C(\varphi \to \chi_1 \vee \cdots \vee \chi_n) \vdash \mathsf{B}^D_C(\chi_1 \vee \cdots \vee \chi_n).$$

The following formula is an instance of the Strict Conditional axiom $\mathsf{K}_C(\varphi \to \chi_1 \vee \cdots \vee \chi_n) \to (\mathsf{B}^D_C(\chi_1 \vee \cdots \vee \chi_n) \to (\varphi \to \mathsf{B}^D_C\varphi))$. Thus, by the Modus Ponens applied twice,

$$X, \varphi, \mathsf{K}_C(\varphi \to \chi_1 \vee \cdots \vee \chi_n) \vdash \varphi \to \mathsf{B}^D_C\varphi.$$

Then, $X, \varphi, \mathsf{K}_C(\varphi \to \chi_1 \vee \cdots \vee \chi_n) \vdash \mathsf{B}^D_C\varphi$ by the Modus Ponens. Thus, $X, \mathsf{K}_C(\varphi \to \chi_1 \vee \cdots \vee \chi_n) \vdash \varphi \to \mathsf{B}^D_C\varphi$ by the deduction lemma. Hence,

$$\{\mathsf{K}_C\overline{\mathsf{K}}_{E_i}\mathsf{B}^{F_i}_{E_i}\chi_i\}_{i=1}^n, \mathsf{K}_C\mathsf{K}_C(\varphi \to \chi_1 \vee \cdots \vee \chi_n) \vdash \mathsf{K}_C(\varphi \to \mathsf{B}^D_C\varphi)$$

by Lemma 3.1 and the definition of set $X$. Then,

$$\{\mathsf{K}_{E_i}\overline{\mathsf{K}}_{E_i}\mathsf{B}^{F_i}_{E_i}\chi_i\}_{i=1}^n, \mathsf{K}_C\mathsf{K}_C(\varphi \to \chi_1 \vee \cdots \vee \chi_n) \vdash \mathsf{K}_C(\varphi \to \mathsf{B}^D_C\varphi)$$

by the Monotonicity axiom, the Modus Ponens inference rule, and the assumption $E_1, \dots, E_n \subseteq C$. Thus,

$$\{\overline{\mathsf{K}}_{E_i}\mathsf{B}^{F_i}_{E_i}\chi_i\}_{i=1}^n, \mathsf{K}_C\mathsf{K}_C(\varphi \to \chi_1 \vee \cdots \vee \chi_n) \vdash \mathsf{K}_C(\varphi \to \mathsf{B}^D_C\varphi)$$

by the definition of modality $\overline{\mathsf{K}}$, the Negative Introspection axiom, and the Modus Ponens rule. Therefore, by Lemma 3.2 and the Modus Ponens inference rule, the statement of the lemma is true. □

## 5 SOUNDNESS

The soundness of the Truth, the Distributivity, the Negative Introspection, the Monotonicity, and the None to Blame axioms is straightforward. Below we prove the soundness of the Joint Responsibility, the Strict Conditional, and the Introspection of Blameworthiness axioms as separate lemmas.

Lemma 5.1. *If $D \cap F = \varnothing$, $(\alpha, \delta, \omega) \Vdash \overline{\mathsf{K}}_C\mathsf{B}^D_C\varphi$, $(\alpha, \delta, \omega) \Vdash \overline{\mathsf{K}}_E\mathsf{B}^F_E\psi$, and $(\alpha, \delta, \omega) \Vdash \varphi \vee \psi$, then $(\alpha, \delta, \omega) \Vdash \mathsf{B}^{D\cup F}_{C\cup E}(\varphi \vee \psi)$.*

Proof. By Definition 2.2 and the definition of modality $\overline{\mathsf{K}}$, assumption $(\alpha, \delta, \omega) \Vdash \overline{\mathsf{K}}_C\mathsf{B}^D_C\varphi$ implies that there is a play $(\alpha_1, \delta_1, \omega_1)$ such that $\alpha \sim_C \alpha_1$ and $(\alpha_1, \delta_1, \omega_1) \Vdash \mathsf{B}^D_C\varphi$. Thus, again by Definition 2.2, there is an action profile $s_1 \in \Delta^D$ such that for each play $(\alpha', \delta', \omega') \in P$, if $\alpha_1 \sim_C \alpha'$ and $s_1 =_D \delta'$, then $(\alpha', \delta', \omega') \nVdash \varphi$. Recall that $\alpha \sim_C \alpha_1$. Thus, for each play $(\alpha', \delta', \omega') \in P$,

$$\alpha \sim_C \alpha' \wedge s_1 =_D \delta' \to (\alpha', \delta', \omega') \nVdash \varphi. \tag{4}$$

Similarly, assumption $(\alpha, \delta, \omega) \Vdash \overline{\mathsf{K}}_E\mathsf{B}^F_E\psi$ implies that there is a profile $s_2 \in \Delta^F$ such that for each play $(\alpha', \delta', \omega') \in P$,

$$\alpha \sim_E \alpha' \wedge s_2 =_F \delta' \to (\alpha', \delta', \omega') \nVdash \psi. \tag{5}$$

Let $s \in \Delta^{D\cup F}$ be the action profile:

$$s(a) = \begin{cases} s_1(a), & \text{if } a \in D, \\ s_2(a), & \text{if } a \in F. \end{cases} \tag{6}$$

Action profile $s$ is well-defined because $D \cap F = \varnothing$. Statements (4), (5), and (6) by Definition 2.2 imply that for each play $(\alpha', \delta', \omega') \in P$ if $\alpha \sim_{C\cup E} \alpha'$ and $s =_{D\cup F} \delta'$, then $(\alpha', \delta', \omega') \nVdash \varphi \vee \psi$. Recall that $(\alpha, \delta, \omega) \Vdash \varphi \vee \psi$. Therefore, $(\alpha, \delta, \omega) \Vdash \mathsf{B}^{D\cup F}_{C\cup E}(\varphi \vee \psi)$ by Definition 2.2. □

Lemma 5.2. *If $(\alpha, \delta, \omega) \Vdash \mathsf{K}_C(\varphi \to \psi)$, $(\alpha, \delta, \omega) \Vdash \mathsf{B}^D_C\psi$, and $(\alpha, \delta, \omega) \Vdash \varphi$, then $(\alpha, \delta, \omega) \Vdash \mathsf{B}^D_C\varphi$.*

Proof. By Definition 2.2, 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 2.2, assumption $(\alpha, \delta, \omega) \Vdash \mathsf{B}^D_C\psi$ implies that there is an action profile $s \in \Delta^D$ such that for each play $(\alpha', \delta', \omega') \in P$, if $\alpha \sim_C \alpha'$ and $s =_D \delta'$, then $(\alpha', \delta', \omega') \nVdash \psi$.

Hence, for each play $(\alpha', \delta', \omega') \in P$, if $\alpha \sim_C \alpha'$ and $s =_D \delta'$, then $(\alpha', \delta', \omega') \nVdash \varphi$. Therefore, $(\alpha, \delta, \omega) \Vdash \mathsf{B}^D_C\varphi$ by Definition 2.2 and the assumption $(\alpha, \delta, \omega) \Vdash \varphi$ of the lemma. □

Lemma 5.3. *If $(\alpha, \delta, \omega) \Vdash \mathsf{B}^D_C\varphi$, then $(\alpha, \delta, \omega) \Vdash \mathsf{K}_C(\varphi \to \mathsf{B}^D_C\varphi)$.*

Proof. By Definition 2.2, assumption $(\alpha, \delta, \omega) \Vdash \mathsf{B}^D_C\varphi$ implies that there is an action profile $s \in \Delta^D$ such that for each play $(\alpha', \delta', \omega') \in P$, if $\alpha \sim_C \alpha'$ and $s =_D \delta'$, 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 2.2, it suffices to show that $(\alpha', \delta', \omega') \Vdash \mathsf{B}^D_C\varphi$.

Consider any play $(\alpha'', \delta'', \omega'') \in P$ such that $\alpha' \sim_C \alpha''$ and $s =_D \delta''$. 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, $(\alpha', \delta', \omega') \Vdash \mathsf{B}^D_C\varphi$ by Definition 2.2 and the assumption $(\alpha', \delta', \omega') \Vdash \varphi$. □

## 6 COMPLETENESS

The standard proof of the completeness for individual knowledge modality $\mathsf{K}_a$ defines states as maximal consistent sets [6]. Two such sets are indistinguishable to an agent $a$ if these sets have the same $\mathsf{K}_a$-formulae. This construction does *not* work for distributed knowledge because if two sets share $\mathsf{K}_a$-formulae and $\mathsf{K}_b$-formulae, they do not necessarily have to share $\mathsf{K}_{a,b}$-formulae. To overcome this issue, we use the Tree of Knowledge construction, similar to the one in [14]. An important change to this construction proposed in the current paper is placing elements of a set $\mathcal{B}$ on the edges of the tree. This change is significant for the proof of Lemma 6.13.

Let $\mathcal{B}$ be an arbitrary set of cardinality larger than that of the set $\mathcal{A}$. Next, for each maximal consistent set of formulae $X_0$, we define the canonical game $G(X_0) = (I, \{\sim_a\}_{a\in\mathcal{A}}, \Delta, \Omega, P, \pi)$.

*Definition 6.1.* The set of outcomes $\Omega$ consists of all sequences $X_0, (C_1, b_1), X_1, (C_2, b_2), \dots, (C_n, b_n), X_n$, where $n \geq 0$ and for each $i \geq 1$, $X_i$ is a maximal consistent subset of $\Phi$, (i) $C_i \subseteq \mathcal{A}$, (ii) $b_i \in \mathcal{B}$, and (iii) $\{\varphi \mid \mathsf{K}_{C_i}\varphi \in X_{i-1}\} \subseteq X_i$.

If $x$ is a nonempty sequence $x_1, \dots, x_n$ and $y$ is an element, then by $x :: y$ and $hd(x)$ we mean sequence $x_1, \dots, x_n, y$ and element $x_n$ respectively.

We say that outcomes $w, u \in \Omega$ are *adjacent* if there are coalition $C$, element $b \in \mathcal{B}$, and maximal consistent set $X$ such that $w = u :: (C, b) :: X$. The adjacency relation forms a tree structure on set $\Omega$, see Figure 2. We call it *the Tree of Knowledge*. We say that edge $(w, u)$ is *labeled* with each agent in coalition $C$ and is *marked* with element $b$. Although vertices of the tree are sequences, it is convenient to think about the maximal consistent set $hd(\omega)$, not a sequence $\omega$, being a vertex of the tree.

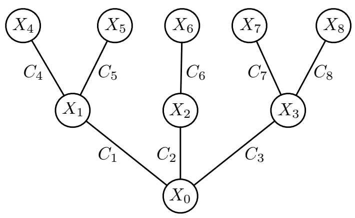


**Figure 2: A Fragment of the Tree of Knowledge.**

*Definition 6.2.* For any outcome $\omega \in \Omega$, let $Tree(\omega)$ be the set of all $\omega' \in \Omega$ such that sequence $\omega$ is a prefix of sequence $\omega'$.

Note that $Tree(\omega)$ is a subtree of the Tree of Knowledge rooted at vertex $\omega$, see Figure 2.

*Definition 6.3.* For any two 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 nodes $\omega$ and $\omega'$ are labeled with agent $a$.

Lemma 6.4. *Relation $\sim_a$ is an equivalence relation on $\Omega$.* □

Lemma 6.5. *$\mathsf{K}_C\varphi \in hd(\omega)$ iff $\mathsf{K}_C\varphi \in hd(\omega')$, if $\omega \sim_C \omega'$.*

Proof. By Definition 6.3, assumption $\omega \sim_C \omega'$ implies that all edges along the unique path between nodes $\omega$ and $\omega'$ are labeled with all agents of coalition $C$. Thus, it suffices to prove the statement of the lemma for any two adjacent vertices along this path. Let $\omega' = \omega :: (D, b) :: X$. Note that $C \subseteq D$ because edge $(\omega, \omega')$ is labeled with all agents in coalition $C$. We start by proving the first part of the lemma.

($\Rightarrow$) Suppose $\mathsf{K}_C\varphi \in hd(\omega)$. Thus, $hd(\omega) \vdash \mathsf{K}_C\mathsf{K}_C\varphi$ by Lemma 3.2. Hence, $hd(\omega) \vdash \mathsf{K}_D\mathsf{K}_C\varphi$ by the Monotonicity axiom. Thus, $\mathsf{K}_D\mathsf{K}_C\varphi \in hd(\omega)$ because set $hd(\omega)$ is maximal. Therefore, $\mathsf{K}_C\varphi \in X = hd(\omega')$ by Definition 6.1.

($\Leftarrow$) Assume $\mathsf{K}_C\varphi \notin hd(\omega)$. Thus, $\neg\mathsf{K}_C\varphi \in hd(\omega)$ by the maximality of the set $hd(\omega)$. Hence, $hd(\omega) \vdash \mathsf{K}_C\neg\mathsf{K}_C\varphi$ by the Negative Introspection axiom. Then, $hd(\omega) \vdash \mathsf{K}_D\neg\mathsf{K}_C\varphi$ by the Monotonicity axiom. Thus, $\mathsf{K}_D\neg\mathsf{K}_C\varphi \in hd(\omega)$ by the maximality of set $hd(\omega)$. Then, $\neg\mathsf{K}_C\varphi \in X = hd(\omega')$ by Definition 6.1. Therefore, $\mathsf{K}_C\varphi \notin hd(\omega')$ because set $hd(\omega')$ is consistent. □

Corollary 6.6. *If $\omega \sim_C \omega'$, then $\overline{\mathsf{K}}_C\varphi \in hd(\omega)$ iff $\overline{\mathsf{K}}_C\varphi \in hd(\omega')$.*

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

*Definition 6.7.* $I = \Omega/\sim_{\mathcal{A}}$.

Lemma 6.8. *Relation $\sim_C$ is well-defined on set $I$.*

Proof. Consider outcomes $\omega_1, \omega_2, \omega_1'$, and $\omega_2'$ where $\omega_1 \sim_C \omega_2$, $\omega_1 \sim_{\mathcal{A}} \omega_1'$, and $\omega_2 \sim_{\mathcal{A}} \omega_2'$. It suffices to show $\omega_1' \sim_C \omega_2'$. Indeed, the assumptions $\omega_1 \sim_{\mathcal{A}} \omega_1'$ and $\omega_2 \sim_{\mathcal{A}} \omega_2'$ imply $\omega_1 \sim_C \omega_1'$ and $\omega_2 \sim_C \omega_2'$. Thus, $\omega_1' \sim_C \omega_2'$ because $\sim_C$ is an equivalence relation. □

Corollary 6.9. *$\alpha \sim_C \alpha'$ iff $\omega \sim_C \omega'$, for any states $\alpha, \alpha' \in I$, any outcomes $\omega \in \alpha$ and $\omega' \in \alpha'$, and any $C \subseteq \mathcal{A}$.*

In [15], the domain of actions $\Delta$ of the canonical game is the set $\Phi$ of all formulae. Informally, if an agent employs action $\varphi$, then she *vetoes* formula $\varphi$. The set $P$ specifies under which conditions the veto takes place. Here, we modify this construction by requiring the agent, while vetoing formula $\varphi$, to specify a coalition $C$ and an outcome $\omega$. The veto will take effect only if coalition $C$ cannot distinguish the outcome $\omega$ from the current outcome. One can think about this construction as requiring the veto ballot to be signed by a key only known, distributively, to coalition $C$. This way only coalition $C$ knows how the agent must vote.

*Definition 6.10.* $\Delta = \{(\varphi, C, \omega) \mid \varphi \in \Phi, C \subseteq \mathcal{A}, \omega \in \Omega\}$.

*Definition 6.11.* The set $P \subseteq I \times \Delta^{\mathcal{A}} \times \Omega$ consists of all triples $(\alpha, \delta, u)$ such that (i) $u \in \alpha$, and (ii) for any outcome $v$ and any formula $\overline{\mathsf{K}}_C\mathsf{B}^D_C\psi \in hd(v)$, if $\delta(a) = (\psi, C, v)$ for each agent $a \in D$ and $u \sim_C v$, then $\neg\psi \in hd(u)$.

*Definition 6.12.* $\pi(p) = \{(\alpha, \delta, \omega) \in P \mid p \in hd(\omega)\}$.

This concludes the definition of the canonical game $G(X_0)$. In Lemma 6.15, we show that this game satisfies the requirement of item (5) from Definition 2.1. Namely, for each $\alpha \in I$ and each complete action profile $\delta \in \Delta^{\mathcal{A}}$, there is at least one $\omega \in \Omega$ such that $(\alpha, \delta, \omega) \in P$.

As usual, the completeness follows from the induction (or "truth") Lemma 6.17. To prove this lemma we first need to establish a few auxiliary properties of game $G(X_0)$.

Lemma 6.13. *For any play $(\alpha, \delta, \omega) \in P$ of game $G(X_0)$, any formula $\neg(\varphi \to \mathsf{B}^D_C\varphi) \in hd(\omega)$, and any profile $s \in \Delta^D$, there is a play $(\alpha', \delta', \omega') \in P$ such that $\alpha \sim_C \alpha'$, $s =_D \delta'$, and $\varphi \in hd(\omega')$.*

Proof. Let the complete action profile $\delta'$ be defined as:

$$\delta'(a) = \begin{cases} s(a), & \text{if } a \in D, \\ (\bot, \varnothing, \omega), & \text{otherwise.} \end{cases} \tag{7}$$

Then, $s =_D \delta'$. Consider the following set of formulae:

$$\begin{aligned} X &= \{\varphi\} \cup \{\psi \mid \mathsf{K}_C\psi \in hd(\omega)\} \\ &\cup \{\neg\chi \mid \overline{\mathsf{K}}_E\mathsf{B}^F_E\chi \in hd(v), E \subseteq C, F \subseteq D, \\ &\qquad \forall a \in F(\delta'(a) = (\chi, E, v)), \omega \sim_E v\}. \end{aligned}$$

Claim 1. *Set $X$ is consistent.*

*Proof of Claim.* Suppose the opposite. Thus, there are formulae $\overline{\mathsf{K}}_{E_1}\mathsf{B}^{F_1}_{E_1}\chi_1, \dots, \overline{\mathsf{K}}_{E_n}\mathsf{B}^{F_n}_{E_n}\chi_n$, outcomes $v_1, \dots, v_n \in \Omega$,

$$\text{and formulae} \quad \mathsf{K}_C\psi_1, \dots, \mathsf{K}_C\psi_m \in hd(\omega), \tag{8}$$
$$\text{such that} \quad \overline{\mathsf{K}}_{E_i}\mathsf{B}^{F_i}_{E_i}\chi_i \in hd(v_i)\ \forall i \le n, \tag{9}$$
$$E_1, \dots, E_n \subseteq C, \quad F_1, \dots, F_n \subseteq D, \tag{10}$$
$$\delta'(a) = (\chi_i, E_i, v_i)\ \forall i \le n\ \forall a \in F_i, \tag{11}$$
$$\omega \sim_{E_i} v_i\ \forall i \le n, \tag{12}$$
$$\text{and} \quad \psi_1, \dots, \psi_m, \neg\chi_1, \dots, \neg\chi_n \vdash \neg\varphi. \tag{13}$$

Without loss of generality, we assume that formulae $\chi_1, \dots, \chi_n$ are distinct. Thus, assumption (11) implies that $F_1, \dots, F_n$ are pairwise disjoint. Assumption (13) implies

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

by the propositional reasoning. Then,

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

by Lemma 3.1. Hence, by assumption (8),

$$hd(\omega) \vdash \mathsf{K}_C(\varphi \to \chi_1 \vee \dots \vee \chi_n).$$

At the same time, $\overline{\mathsf{K}}_{E_1}\mathsf{B}^{F_1}_{E_1}\chi_1, \dots, \overline{\mathsf{K}}_{E_n}\mathsf{B}^{F_n}_{E_n}\chi_n \in hd(\omega)$ by assumption (9), assumption (12), and Corollary 6.6. Thus, $hd(\omega) \vdash \mathsf{K}_C(\varphi \to \mathsf{B}^D_C\varphi)$ by Lemma 4.5, assumption (10), and the assumption that sets $F_1, \dots, F_n$ are pairwise disjoint. Hence, by the Truth axiom, $hd(\omega) \vdash \varphi \to \mathsf{B}^D_C\varphi$, which contradicts the assumption $\neg(\varphi \to \mathsf{B}^D_C\varphi) \in hd(\omega)$ of the lemma because set $hd(\omega)$ is consistent. Thus, $X$ is consistent. ⊠

Let $X'$ be any maximal consistent extension of set $X$ and $\omega'_b$ be the sequence $\omega :: (C, b) :: X'$ for each element $b \in \mathcal{B}$. Then, $\omega'_b \in \Omega$ for each element $b \in \mathcal{B}$ by Definition 6.1 and the choice of sets $X$ and $X'$. Also $\varphi \in X \subseteq hd(\omega'_b)$ for each $b \in \mathcal{B}$ by the choice of sets $X$ and $X'$.

Note that family $\{Tree(\omega'_b)\}_{b\in\mathcal{B}}$ consists of pair-wise disjoint sets. This family has the same cardinality as set $\mathcal{B}$. Let

$$V = \{v \in \Omega \mid \delta'(a) = (\psi, E, v), a \in \mathcal{A}, \psi \in \Phi, E \subseteq \mathcal{A}\}.$$

The cardinality of $V$ is at most the cardinality of set $\mathcal{A}$. By the choice of set $\mathcal{B}$, its cardinality is larger than the cardinality of set $\mathcal{A}$. Thus, there exists a set $Tree(\omega'_{b_0})$ in family $\{Tree(\omega'_b)\}_{b\in\mathcal{B}}$ disjoint with set $V$:

$$Tree(\omega'_{b_0}) \cap V = \varnothing. \tag{14}$$

Let $\omega'$ be the outcome $\omega'_{b_0}$.

Claim 2. *If $\omega' \sim_E v$ for some $v \in V$, then $E \subseteq C$.*

*Proof of Claim.* Consider any agent $a \in E$. By Definition 6.3, assumption $\omega' \sim_E v$ implies that each edge along the unique path connecting vertex $\omega$ with vertex $v$ is labeled with agent $a$. At the same time, $v \notin Tree(\omega')$ by statement (14) and because $\omega' = \omega'_{b_0}$. Thus, the path between vertex $\omega'$ and vertex $v$ must go through vertex $\omega$, see Figure 3. Hence, this path must contain edge $(\omega', \omega)$. Since all edges along this path are labeled with agent $a$ and edge $(\omega', \omega)$ is labeled with agents from set $C$, it follows that $a \in C$. ⊠

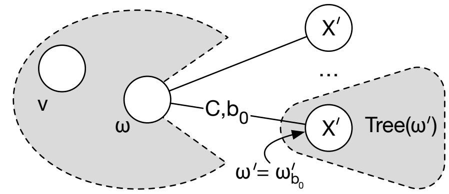


**Figure 3: Towards the Proof of Claim 2.**

Let initial state $\alpha'$ be the equivalence class of outcome $\omega'$ with respect to the equivalence relation $\sim_{\mathcal{A}}$. Note that $\omega \sim_C \omega'$ by Definition 6.1 because $\omega' = \omega :: (C, b_0) :: X'$. Therefore, $\alpha \sim_C \alpha'$ by Corollary 6.9.

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

*Proof of Claim.* First, note that $\omega' \in \alpha'$ because initial state $\alpha'$ is the equivalence class of outcome $\omega'$. Next, consider an outcome $v \in \Omega$

$$\text{and a formula} \quad \overline{\mathsf{K}}_E\mathsf{B}^F_E\chi \in hd(v), \tag{15}$$
$$\text{such that} \quad \omega' \sim_E v, \tag{16}$$
$$\text{and} \quad \forall a \in F\,(\delta'(a) = (\chi, E, v)). \tag{17}$$

By Definition 6.11, it suffices to show that $\neg\chi \in hd(\omega')$.

**Case I:** $F = \varnothing$. Then, $\neg\mathsf{B}^F_E\chi$ is an instance of the None to Act axiom. Thus, $\vdash \mathsf{K}_E\neg\mathsf{B}^F_E\chi$ by the Necessitation inference rule. Hence, $\neg\mathsf{K}_E\neg\mathsf{B}^F_E\chi \notin hd(v)$ by the consistency of the set $hd(v)$, which contradicts the assumption (15) and the definition of modality $\overline{\mathsf{K}}$.

**Case II:** $\varnothing \ne F \subseteq D$. Thus, there exists an agent $a \in F$. Note that $\delta'(a) = (\chi, E, v)$ by assumption (17). Hence, $v \in V$ by the definition of set $V$. Thus, $E \subseteq C$ by Claim 2 and assumption (16). Then, $\neg\chi \in X$ by the definition of set $X$, the assumption of the case that $F \subseteq D$, assumption (15), assumption (16), and assumption (17). Therefore, $\neg\chi \in hd(\omega')$ because $X \subseteq X' = hd(\omega'_{b_0}) = hd(\omega')$ by the choice of set $X'$, set of sequences $\{\omega'_b\}_{b\in\mathcal{B}}$, and sequence $\omega'$.

**Case III:** $F \nsubseteq D$. Consider any agent $a \in F \setminus D$. Thus, $\delta'(a) = (\bot, \varnothing, \omega)$ by equation (7). Thus, $\chi \equiv \bot$ by statement (17) and the assumption $a \in F$. 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 6.14. *For any outcome $\omega \in \Omega$, there is a state $\alpha \in I$ and a complete profile $\delta \in \Delta^{\mathcal{A}}$ such that $(\alpha, \delta, \omega) \in P$.*

Proof. Let initial state $\alpha$ be the equivalence class of outcome $\omega$ with respect to the equivalence relation $\sim_{\mathcal{A}}$. Thus, $\omega \in \alpha$. Let $\delta$ be the complete profile such that $\delta(a) = (\bot, \varnothing, \omega)$ for each $a \in \mathcal{A}$. To prove $(\alpha, \delta, \omega) \in P$, consider any outcome $v \in \Omega$, any formula $\overline{\mathsf{K}}_C\mathsf{B}^D_C\chi \in hd(v)$ such that

$$\forall a \in D\,(\delta(a) = (\chi, C, v)). \tag{18}$$

By Definition 6.11, it suffices to show that $\neg\chi \in hd(\omega)$.

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

**Case II**: $D \neq \varnothing$. Then, there is an agent $a \in D$. Thus, $\delta(a) = (\chi, C, v)$ by statement (18). Hence, $\chi \equiv \bot$ by the definition of action profile $\delta$. Then, $\neg\chi$ is a tautology. Therefore, $\neg\chi \in hd(\omega)$ by the maximality of set $hd(\omega)$. □

Lemma 6.15. *For each $\alpha \in I$ and each complete action profile $\delta \in \Delta^{\mathcal{A}}$, there is at least one outcome $\omega \in \Omega$ such that $(\alpha, \delta, \omega) \in P$.*

Proof. By Definition 6.7, initial state $\alpha$ is an equivalence class. Since each equivalence class is not empty, there must exist an outcome $\omega_0 \in \Omega$ such that $\omega_0 \in \alpha$. By Lemma 6.14, there is an initial state $\alpha_0 \in I$ and a complete action profile $\delta_0 \in \Delta^{\mathcal{A}}$ such that $(\alpha_0, \delta_0, \omega_0) \in P$. Then, $\omega_0 \in \alpha_0$ by Definition 6.11. Hence, $\omega_0$ belongs to both equivalence classes $\alpha$ and $\alpha_0$. Thus, $\alpha = \alpha_0$. Therefore, $(\alpha, \delta_0, \omega_0) \in P$. □

Lemma 6.16. *For any play $(\alpha, \delta, \omega) \in P$ 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 3.1. 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.

Recall that set $\mathcal{B}$ has larger cardinality than set $\mathcal{A}$. Thus, there is at least one $b \in \mathcal{B}$. Let set $X'$ be any maximal consistent extension of set $X$ and $\omega'$ be the sequence $\omega :: (C, b) :: X'$. Note that $\omega' \in \Omega$ by Definition 6.1 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 6.14, there is an initial state $\alpha' \in I$ and a complete action profile $\delta'$ such that $(\alpha', \delta', \omega') \in P$. Note that $\omega \sim_C \omega'$ by Definition 6.3 and the choice of sequence $\omega'$. Thus, $\alpha \sim_C \alpha'$ by Corollary 6.9. □

Lemma 6.17. $(\alpha, \delta, \omega) \Vdash \varphi$ *iff* $\varphi \in hd(\omega)$.

Proof. We prove the lemma by induction on the complexity of formula $\varphi$. If $\varphi$ is a propositional variable, then the lemma follows from Definition 2.2 and Definition 6.12. If formula $\varphi$ is an implication or a negation, then the required follows from the induction hypothesis and the maximality and the consistency of set $hd(\omega)$ by Definition 2.2. Assume that formula $\varphi$ 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 6.16, 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 2.2.

$(\Leftarrow)$ : 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 6.5. 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 6.9. Therefore, $(\alpha, \delta, \omega) \Vdash \mathsf{K}_C\psi$ by Definition 2.2.

Assume formula $\varphi$ has the form $\mathsf{B}^D_C\psi$.

$(\Rightarrow)$ : Suppose $\mathsf{B}^D_C\psi \notin hd(\omega)$.

**Case I**: $\psi \notin hd(\omega)$. Then, $(\alpha, \delta, \omega) \nVdash \psi$ by the induction hypothesis. Thus, $(\alpha, \delta, \omega) \nVdash \mathsf{B}^D_C\psi$ by Definition 2.2.

**Case II**: $\psi \in hd(\omega)$. Let us show that $\psi \to \mathsf{B}^D_C\psi \notin hd(\omega)$. Indeed, if $\psi \to \mathsf{B}^D_C\psi \in hd(\omega)$, then $hd(\omega) \vdash \mathsf{B}^D_C\psi$ by the Modus Ponens rule. Thus, $\mathsf{B}^D_C\psi \in hd(\omega)$ by the maximality of set $hd(\omega)$, which contradicts the assumption above.

Since set $hd(\omega)$ is maximal, statement $\psi \to \mathsf{B}^D_C\psi \notin hd(\omega)$ implies that $\neg(\psi \to \mathsf{B}^D_C\psi) \in hd(\omega)$. Hence, by Lemma 6.13, for any action profile $s \in \Delta^D$, there is a play $(\alpha', \delta', \omega')$ such that $\alpha \sim_C \alpha'$, $s =_D \delta'$, and $\psi \in hd(\omega')$. Thus, by the induction hypothesis, for any action profile $s \in \Delta^D$, there is a play $(\alpha', \delta', \omega')$ such that $\alpha \sim_C \alpha'$, $s =_D \delta'$, and $(\alpha', \delta', \omega') \Vdash \psi$. Therefore, $(\alpha, \delta, \omega) \nVdash \mathsf{B}^D_C\psi$ by Definition 2.2.

$(\Leftarrow)$ : Let $\mathsf{B}^D_C\psi \in hd(\omega)$. 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 $s \in \Delta^D$ be the action profile of coalition $D$ such that $s(a) = (\psi, C, \omega)$ for each agent $a \in D$. Consider any play $(\alpha', \delta', \omega') \in P$ such that $\alpha \sim_C \alpha'$ and $s =_D \delta'$. By Definition 2.2, it suffices to show that $(\alpha', \delta', \omega') \nVdash \psi$.

Assumption $\mathsf{B}^D_C\psi \in hd(\omega)$ implies $hd(\omega) \nvdash \neg\mathsf{B}^D_C\psi$ because set $hd(\omega)$ is consistent. Thus, $hd(\omega) \nvdash \mathsf{K}_C\neg\mathsf{B}^D_C\psi$ by the contraposition of the Truth axiom. Hence, $\neg\mathsf{K}_C\neg\mathsf{B}^D_C\psi \in hd(\omega)$ by the maximality of $hd(\omega)$. Then, $\overline{\mathsf{K}}_C\mathsf{B}^D_C\psi \in hd(\omega)$ by the definition of modality $\overline{\mathsf{K}}$. Recall that $s(a) = (\psi, C, \omega)$ for each agent $a \in D$ by the choice of the action profile $s$. Also, $s =_D \delta'$ by the choice of the play $(\alpha', \delta', \omega')$. Hence, $\delta'(a) = (\psi, C, \omega)$ for each agent $a \in D$. Thus, $\neg\psi \in hd(\omega')$ by Definition 6.11 and because $\overline{\mathsf{K}}_C\mathsf{B}^D_C\psi \in hd(\omega)$ and $(\alpha', \delta', \omega') \in P$. Then, $\psi \notin hd(\omega')$ by the consistency of set $hd(\omega')$. Therefore, $(\alpha', \delta', \omega') \nVdash \psi$ by the induction hypothesis. □

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

Theorem 6.18. *If $X \nvdash \varphi$, then there is a game, and a play $(\alpha, \delta, \omega)$ 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 $\omega_0$ be the single-element sequence $X_0$. Note that $\omega_0 \in \Omega$ by Definition 6.1. By Lemma 6.14, 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 6.17 and the choice of set $X_0$. Thus, $(\alpha, \delta, \omega_0) \nVdash \varphi$ by Definition 2.2. □

## 7 CONCLUSION

In this paper, we proposed a formal definition of the second-order blameworthiness or duty to warn in the setting of strategic games. Our main technical result is a sound and complete logical system that describes the interplay between the second-order blameworthiness and the distributed knowledge modalities.

## REFERENCES


[1] Thomas Ågotnes and Natasha Alechina. 2019. Coalition Logic with Individual, Distributed and Common Knowledge. *Journal of Logic and Computation* 29 (11 2019), 1041–1069. Issue 7. https://doi.org/10.1093/logcom/exv085

[2] Nina Renata Aron. 2017. The murder of this 20-year-old Berkeley coed changed the laws around psychology forever. *Timeline* (26 September 26, 2017). https://timeline.com/tanya-tarasoff-notify-law-7d43951cb004

[3] 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)*.

[4] 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

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

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

[7] 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.

[8] Harry G Frankfurt. 1969. Alternate possibilities and moral responsibility. *The Journal of Philosophy* 66, 23 (1969), 829–839. https://doi.org/10.2307/2023833

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

[10] David Lewis. 2013. *Counterfactuals*. John Wiley & Sons.

[11] 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.

[12] 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.

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

[14] 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

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

[16] Pavel Naumov and Jia Tao. 2019. Knowing-how under uncertainty. *Artificial Intelligence* 276 (2019), 41 – 56. https://doi.org/10.1016/j.artint.2019.06.007

[17] 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

[18] J. Tobriner. 1976. Majority Opinion. (1976). S.F. No. 23042. Supreme Court of California. July 1, 1976.

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

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