

# Counterfactual and Seeing-to-it Responsibilities in Strategic Games

Pavel Naumov<sup>a</sup>, Jia Tao<sup>b,\*</sup>

<sup>a</sup>*University of Southampton, Southampton, United Kingdom*

<sup>b</sup>*Lafayette College, Easton, PA, United States*

---

## Abstract

The article studies two forms of responsibility in the setting of strategic games with imperfect information. They are referred to as seeing-to-it responsibility and counterfactual responsibility. It shows that counterfactual responsibility is definable through seeing-to-it, but not the other way around. The article also proposes a sound and complete bimodal logical system that describes the interplay between the seeing-to-it modality and the individual ex ante knowledge modality.

*Keywords:* responsibility, undefinability, axiomatization, completeness

---

## 1. Introduction

In this article, we study formal semantics of responsibility. In the literature, there have been two different approaches to defining responsibility.

The first approach is based on what became known as Frankfurt's principle of alternate possibilities: "a person is morally responsible for what he has done only if he could have done otherwise" [1]. The principle of alternative possibilities is widely discussed in the literature [2]. Although Frankfurt and many others agree that this principle has many exceptions and limitations, the principle is often taken as a starting point in philosophical discussions of responsibility. This principle, sometimes referred to as "counterfactual possibility" [3], is also used to define causality [4, 5, 6]. For the sake of clarity, in

---

\*Corresponding author

*Email addresses:* p.naumov@soton.ac.uk (Pavel Naumov), taoj@lafayette.edu (Jia Tao)

this article, we refer to all versions of responsibility based on the principle of alternative possibilities as *counterfactual responsibility*. Formal logical systems for reasoning about this form of responsibility in strategic and security games are proposed in [7] and [8] respectively.

The other approach is to hold a person responsible for the outcome if the person makes it unavoidable that the outcome happens. In this article, we refer to this approach as responsibility for *seeing-to-it*. This approach to responsibility has been extensively studied in STIT (“seeing to it that”) logic [9, 10, 11, 12, 13].

### 1.1. Responsibility in Strategic Games

The difference between the two forms of responsibility could be illustrated using two strategic games depicted in Figure 1. We refer to these as the left

	d <sub>1</sub>	d <sub>2</sub>	d <sub>3</sub>
m <sub>1</sub>			
m <sub>2</sub>			😭 /
m <sub>3</sub>	😭	😭	

	d <sub>1</sub>	d <sub>2</sub>	d <sub>3</sub>
m <sub>1</sub>		😭 /	
m <sub>2</sub>	😭	😭	😭
m <sub>3</sub>	😭		

Figure 1: Under action profile  $(m_2, d_3)$  in the left game, if baby cries, the mom is *counterfactually responsible* for baby crying. Under the same action profile in the right game, she *sees to it* that the baby cries.

game and the right game. In these games, the agents mom and dad are trying to prevent their baby from crying. In both games, each parent has three strategies:  $m_1$ ,  $m_2$ , and  $m_3$  for mom and  $d_1$ ,  $d_2$ , and  $d_3$  for dad. The cells of the tables represent action profiles. The crying emoji marks action profiles under which the baby cries. In this article, we consider nondeterministic games that might have multiple outcomes for the same action profile. If an action profile might result in multiple outcomes, we further split the cell into triangles representing these outcomes. For example, in the left game, under action profile  $(m_2, d_3)$  there might be two possible outcomes. Only in one of them does the baby cry.

Consider a situation when parents choose actions  $m_2$  and  $d_3$  in the left game and the baby cries. In this case, according to the principle of alternative

possibilities, the mom is counterfactually responsible for the baby crying because the mom could have prevented it by choosing action  $m_1$ . At the same time, in the right game, under the same profile  $(m_2, d_3)$ , the baby also cries, but the mom is not counterfactually responsible for it because in the right game she has no unilateral action that would prevent the baby from crying.

However, the mom is responsible *for seeing to* the baby’s crying in the right game under action profile  $(m_2, d_3)$ . Indeed, by choosing action  $m_2$  the mom guarantees that the baby cries. On the other hand, she is not responsible for seeing-to-it under profile  $(m_2, d_3)$  in the left game because action  $m_2$  in the left game does not unavoidably lead to the baby crying.

Note that statement  $2 + 2 = 4$  is true no matter what actions the mom and the dad choose. Thus, one can say that the mom sees to it that  $2 + 2 = 4$ . This is the approach taken in many works on STIT logic. However, such an approach is problematic if “seeing to it” is interpreted as a form of responsibility. One can hold an agent responsible for seeing-to-it that something happens only if the agent had an alternative action that does not unavoidably lead to it. Horty and Belnap refer to seeing-to-it in the presence of such an alternative action as seeing-to-it “deliberately” [11]. Since the focus of our article is on responsibility, we include the existence of the alternative action in our definition of seeing-to-it. In the right game from Figure 1, under action profile  $(m_2, d_3)$  such an alternative action of mom not unavoidably leading to baby crying is, for example, action  $m_3$ .

### 1.2. Examples from the Literature

To further illustrate the two forms of responsibility, we turn to three examples from the literature.

**Example 1.** *“Billy and Suzy throw rocks at a bottle. Suzy throws first, or maybe she throws harder. Her rock arrives first. The bottle shatters. When Billy’s rock gets to where the bottle used to be, there is nothing there but flying shards of glass. Without Suzy’s throw, the impact of Billy’s rock on the intact bottle would have been one of the final steps in the causal chain from Billy’s throw to the shattering of the bottle. But, thanks to Suzy’s preempting throw, that impact never happens.” – D. Lewis [14].*

This situation can be captured by a nondeterministic strategic game between two agents, Billy and Suzy, depicted in Figure 2. Each of them has two








		not throw	throw
not throw			
throw			




Figure 2: Rock throwing game between Billy and Suzy. One of the players chooses a row and the other chooses a column.

actions: “throw” and “not throw”. The bottle *might* get broken if either player decides to throw a rock. The example above refers to the action profile (throw, throw) and the outcome when the bottle is shattered. Note that throwing the rock does *not* unavoidably lead to a shattered bottle. Thus, neither Billy nor Suzy is responsible for *seeing to* the bottle being shattered. At the same time, neither of them is counterfactually responsible because none of them has a unilateral action that would prevent the bottle from being shattered.





**Example 2.** “Suppose that two companies both dump pollutant into the river. Company A dumps 100 kilograms of pollutant; company B dumps 60 kilograms. The fish in the river die. Biologists determine that  $k$  kilograms of pollutant suffice for the fish to die. Which company is the cause of the fish dying if  $k = 120$ , if  $k = 80$ , and if  $k = 50$ ?” – J. Halpern [15].

		not dump	dump
not dump			
dump			

k=120

		not dump	dump
not dump			
dump			

k=80

		not dump	dump
not dump			
dump			

k=50

Figure 3: Three pollution games, corresponding to different values of  $k$ . In each game, Company A selects a row and Company B selects a column.

Cases of  $k = 120, 80, 50$  can be captured by the three strategic games depicted in Figure 3. The outcomes in which fish die are marked with a dead fish

picture. The example refers to the profile (dump, dump). If  $k = 120$ , then neither of the companies sees to the death of the fish, but both of them are counterfactually responsible for the death. If  $k = 80$ , then only Company A sees to the death and it is the only one who is counterfactually responsible for the death. If  $k = 50$ , then both of the companies sees to the death and neither of them is counterfactually responsible for it.

**Example 3.** *“You are the manager of your home country’s team in the International Salsa Competition. Your team consists of Alice, Bob, Chuck, and Dan. In order to compete in the tournament, Alice will need to show up and at least one of her partners. You instruct all of them to come to the tournament. However, as it turns out, none of them show up on the day of the competition.”* – R. Zultan, T. Gerstenberg, D. A. Lagnado [16]

This situation can be modelled as a strategic game between agents Alice, Bob, Chuck, and Dan. Each agent has two actions: “show up” and “not show up”. The example refers to the action profile under which all agents choose action “not show up”. Under the given scenario, Alice sees to it that the team is not competing in the tournament because (i) her action would unavoidably result in the team being disqualified from the tournament and (ii) there is a possibility of not being disqualified if she decides to show up. Neither Bob, Chuck, nor Dan sees to it that the team is not competing in the tournament. None of the four of them is individually counterfactually responsible for the disqualification because neither of them had a strategy to prevent it. At the same time, four of them as a group are counterfactually responsible for the outcome using the definition of group counterfactual responsibility from [7].

### 1.3. Responsibility and Knowledge

Knowledge is an important factor in ascribing responsibility to agents. The connection between responsibility and knowledge has been discussed by philosophers since Aristotle [17]:

*... blame is given only to what is voluntary ... a voluntary act is one which is originated by the doer with knowledge of the particular circumstances of the act.*

– *The Nicomachean Ethics, Book III, Chapters 1-5*

In a legal setting, responsibility is also commonly defined as a combination of knowledge and actions, often referred to as *guilty mind* and *guilty actions*. For example, US Model Penal Code distinguishes five such combinations referred to as strict liability and liability for doing negligently, recklessly, knowingly, and purposefully [18].

If one considers games with imperfect information instead of games with perfect information that we discussed in the previous subsection, then the definitions of both forms of responsibility must be adjusted to incorporate knowledge. In the case of counterfactual responsibility, it is natural to require that the agent is responsible for an outcome when she not only has a strategy to prevent it, but also knows *ex ante* (before the action) what this strategy is [19, 20, 21]. In the case of seeing-to-it, it is natural to require that not only the agent’s action unavoidably leads to the outcome, but the agent also must *interim* (at the time of the action) know this [22].

To illustrate the two definitions of responsibility in the imperfect information setting, consider an execution of a death penalty by shooting. If the execution is administered by a single shooter, then the shooter is responsible for the death of the prisoner under both definitions. Indeed, the shooter is counterfactually responsible because the prisoner is dead after the shooting, and the shooter knows *ex ante* that this could be prevented by not firing the lethal shot. The shooter is also responsible for seeing to the death because the shooter knows *interim* (at the moment the trigger is pulled) that the action will result in death. The shooter also knows that the prisoner might not die if the trigger is not pulled.

Most executions by shooting are performed by a firing squad rather than by a single shooter. If multiple shooters are instructed to fire simultaneously, then no single shooter has a strategy to prevent the death of the prisoner. Thus, *none of them is counterfactually responsible for the death individually*<sup>1</sup>. At the same time, each of the agents knows *interim* that pulling the trigger while aiming at the prisoner will unavoidably result in death, while not shooting leaves a possibility (if all other shooters do not shoot too) for the prisoner not to be killed. Thus, *each member* of the firing squad who pulls the trigger while aiming at the prisoner *is responsible for seeing to the death of the prisoner*.

---

<sup>1</sup>All shooters could be blamed together as a group under a coalitional counterfactual responsibility definition such as, for example, in [19].

In some cases, one or more members of the firing squad are issued a weapon containing a blank cartridge,<sup>2</sup> also known as the “conscience bullet”. The members of the squad are told that one of them has a blank cartridge, but they are not told which one. Because the blank cartridge has no bullet, it gives no recoil. As a result, each shooter knows *ex post* (after the trigger is pulled) which cartridge it was, but not *ex ante* or *interim* (at the moment the trigger is pulled). If one or more members of the firing squad are issued “conscience bullets”, then *none of the squad members are counterfactually responsible* for the death. Indeed, even if only one member of the squad is issued the real bullet, then this member has a strategy to prevent the death, but *the shooter does not know this*. If more than one of them is issued a real bullet, then none of the members has a strategy to prevent the death unilaterally. In the “conscience bullet” setting, *none of the agents is responsible for seeing to the death* of the prisoner because none of them knows *interim* that pulling the trigger while aiming at the prisoner will unavoidably result in the death of the prisoner.

Thus, an execution by a single shooter results in the shooter being responsible for seeing to the death and also responsible for the death counterfactually. If an execution by a firing squad is without a “conscience bullet”, then none of the squad members is responsible counterfactually, but each of them is responsible for seeing to the death. If at least one “conscience bullet” is used, then none of the members is responsible for seeing to the death. Nor is any of them responsible for the death counterfactually.

#### 1.4. Contribution and Outline

In this article, we study the seeing-to-it and the counterfactual forms of responsibility. Instead of investigating these two notions in isolation, we focus on a perhaps deeper question: the interplay between them. One way to address this question is to study the definability of these notions through each other. The other is to capture the universal properties of the interplay in a complete logical system. In this article, we do both. First, we show that the counterfactual form of responsibility can be defined through the seeing-to-it responsibility but not the other way around. Second, we give a sound and complete axiomatisation of the seeing-to-it form of responsibility for the

---

<sup>2</sup>“The officer charged with the execution will . . . Cause eight rifles to be loaded in his presence. Not more than three nor less than one will be loaded with blank ammunition. He will place the rifles at random in the rack provided for that purpose.” [23, p.5]

class of the models that we consider. Multiple axiomatisations of seeing-to-it responsibility for different classes of semantics have been proposed before for the responsibility defined without the requirement that the agent had an alternative action that does not unavoidably lead to the statement being true [24, 25, 26]. We discuss this modality in Section 4.1. The originality of the axiomatic part of our contribution is in treating the seeing-to-it responsibility with that requirement as a single modality. The preliminary version of this paper, without axiomatisation and the proof of completeness, appeared as [27].

The rest of this article is organised as follows. In Section 2, we propose a formal semantics of seeing-to-it responsibility and counterfactual responsibility as modalities in the strategic game setting. In Section 3, we show that the counterfactual responsibility modality is definable through the seeing-to-it modality, but not the other way around. In Section 4, we give a sound and complete axiomatisation of the seeing-to-it modality. Section 5 serves as a conclusion.

## 2. Formal Definitions

### 2.1. Imperfect Information Strategic Games

In this section, we give the formal definition of games with imperfect information used in the rest of the article. In the context of this article, imperfect information refers to the fact that an agent might not know everything about the current situation *before* the agent takes an action. For example, each member of the firing squad does not know whose weapon is loaded with a blank cartridge. As it is common in the field of epistemic logic, we model this type of uncertainty by having multiple *initial states* and an *indistinguishability* relation  $\sim_a$  on these states, specific to each agent  $a$ . The knowledge captured by relation  $\sim_a$  is the *ex ante* knowledge of agent  $a$ . Because names of the agents appear in the language of our logical system, it is convenient to assume that the set of agents is fixed throughout the article. We denote this set by  $\mathcal{A}$ .

The notion of responsibility has been traditionally based on the assumption that agents have *free will*. We model free will by assuming that each agent has a nonempty set of *actions* from which the agent needs to choose one. We treat “abstaining” as one of the possible actions. In our firing squad example, each shooter has only two actions: “shoot” and “abstain”. By  $\Delta_a^\alpha$  we denote the set of all actions available to agent  $a$  in the initial state  $\alpha$ .



If an agent  $a$  cannot distinguish initial states  $\alpha$  and  $\alpha'$  and  $d \in \Delta_a^\alpha \setminus \Delta_a^{\alpha'}$ , then in initial state  $\alpha$  agent  $a$  does not know (is not *aware*) that action  $d$  is available. In this article, we assume that in each initial state each agent is aware of all available actions. We capture this *uniformity* assumption by requiring that  $\Delta_a^\alpha = \Delta_a^{\alpha'}$  for any initial states  $\alpha$  and  $\alpha'$  such that  $\alpha \sim_a \alpha'$ .

As common in game theory, by an *action profile* we mean a function that assigns an action to each agent. Just like we did in our example in Figure 1, we assume that the same action profile might lead to different outcomes. This nondeterminicity can potentially be modelled in three different ways. First, we can add an additional agent, Goddess, who acts simultaneously with the other agents and whose action, together with the actions of the other agents, would uniquely determine the outcome. Second, we can assume that Goddess chooses her action before the other agents, but does not disclose it. This means that Goddess' action would be captured by the initial state. In this case, essentially, the uncertainty of the action is modelled through the uncertainty of the initial state. Finally, we can model nondeterminicity by introducing outcome states, or just *outcomes*, and assuming that the same action profile in the same initial state might lead to different outcome states. Clearly, these three ways to model nondeterminicity are all equivalent, in the sense that any game using one of them can be converted into a game using any other. In our formal definition, we have chosen the third approach because it leads to a somewhat simpler proof of completeness. We denote the set of all outcomes by  $\Omega$ .

To capture “the rules of the game”, we use plays. A play is a *possible* combination of an initial state, an action profile, and an outcome. If all such combinations are possible, then the game has no rules and there is no connection between the initial state, the individual actions of the agents, and the outcome. The outcome of such a game is completely unpredictable. In general, by choosing a set  $P$  of (“valid”) plays, we fix an inter-dependency between initial states, actions, and outcomes. Trying to be as general as possible, we assume very little about set  $P$ . Namely, we only require that for each initial state and each action profile there is at least one possible outcome. This, essentially, excludes the possibility for a game in some initial state to terminate without reaching an outcome. The reason for this exclusion is that ascribing responsibility under plays without an outcome is problematic. Indeed, should terminating a game be viewed as a prevention of an undesired event? Should forcing the termination of a game be viewed as seeing to such an event? Or should the termination be viewed as an even more undesirable

“world apocalypses”?

As usual, the language of our logical system includes a nonempty set of propositional variables. In modal logic, it is common to interpret propositional variables as statements about states. In our case, that would probably mean interpreting propositional variables as statements about outcomes. In this article, we take a more general approach of interpreting propositional variables as statements about plays, rather than just states. For example, a propositional variable can denote the statement “Suzy did not throw the rock, but the bottle is broken” whose first part refers to the action profile and the second to the outcome. As a special case, our propositional variables can also represent statements about just the outcome. Thus, the valuation  $\pi$  of a propositional variable is a subset of  $P$ . We further discuss this in Section 2.4.

**Definition 1.** *A game is a tuple  $(I, \{\sim_a\}_{a \in \mathcal{A}}, \{\Delta_a^\alpha\}_{a \in \mathcal{A}}^{\alpha \in I}, \Omega, P, \pi)$ , where*

1.  $I$  is a (possibly empty) 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.  $\Delta_a^\alpha$  is a nonempty set of **actions** for each agent  $a \in \mathcal{A}$  and each initial state  $\alpha \in I$ , which satisfies the **uniformity** assumption: for each agent  $a \in \mathcal{A}$  and all initial states  $\alpha, \alpha' \in I$ , if  $\alpha \sim_a \alpha'$ , then  $\Delta_a^\alpha = \Delta_a^{\alpha'}$ ,
4.  $\Omega$  is a set of **outcomes**,
5.  $P$  is a set of triples  $(\alpha, \delta, \omega)$ , called **plays**, where  $\alpha \in I$  is an initial state,  $\omega \in \Omega$  is an outcome, and
  - (a) function  $\delta$ , called an **action profile in state**  $\alpha$ , is such that  $\delta(a) \in \Delta_a^\alpha$  for each initial state  $\alpha \in I$  and each agent  $a \in \mathcal{A}$ ,
  - (b) for each initial state  $\alpha \in I$  and each action profile  $\delta$  in state  $\alpha$ , there is at least one outcome  $\omega \in \Omega$  such that  $(\alpha, \delta, \omega) \in P$ ,
6.  $\pi(p)$  is a subset of  $P$  for each propositional variable  $p$ .

Multiple formal frameworks for capturing the interplay between knowledge and ability have been proposed before. Some of them follow the computer science tradition of modelling abilities through actions. Others, in the philosophical tradition of STIT logic, model abilities through either relations or types.

Our Definition 1 follows the action approach. Our notion of game is very similar to a concurrent game structure with imperfect information [28, 29],

an epistemic transition system [30, 31], a game [20, 32], an epistemic coalition model [33], and a concurrent epistemic game structure [34]. Unlike the above models, in Definition 1, we distinguish the set of initial states  $I$  from the set of outcomes  $\Omega$ . This is done for notational convenience only. In particular, we allow  $I$  and  $\Omega$  to be the same set. According to our formal semantics given in Definition 2, nesting of modalities is possible even if sets  $I$  and  $\Omega$  are disjoint. In fact, nesting is used in Section 3.1 to express the modality representing counterfactual responsibility through the modality representing seeing-to-it responsibility.

Compared to works in STIT tradition, our Definition 1 is most similar to Horty and Pacuit’s epistemic stit frames [12]. The initial states in our definition correspond to moments in their semantics. Our plays correspond to indices (a pair consisting of a history and a moment of that history). The actions in our games correspond to types in epistemic stit frames. Relation-based semantics of STIT [35, 36, 37, 26] use an agent-specific relation “acted the same way” on indices. The equivalence classes under such a relation correspond to types in [12] and to actions in our games.

A more abstract class of game models is advanced by Lorini, Longin, and Mayor [22]. Instead of taking initial states, actions, and outcomes as atomic objects, they consider plays as such objects. In addition to the set of plays, the game models have “projection” functions that map each play into the action taken by a given agent on that play as well as epistemic indistinguishability relations on the plays. The game models in the current article can be converted into that more abstract type of game models and vice versa.

## 2.2. Pollution game with imperfect information

In this section, we give an example of a game with imperfect information inspired by Halpern’s pollution example, see Example 2 in Section 1.2. This game has three initial states corresponding to values  $k = 120, 80, 50$ , see Figure 4. We assume that the actual value of the threshold after which the fish die is  $k = 80$ , but neither of the companies knows this. To be more specific, let us assume that Company A cannot distinguish initial states  $k = 120$  and  $k = 80$  while Company B cannot distinguish initial states  $k = 80$  and  $k = 50$ . In Figure 4, we show the indistinguishability relation using dashed lines.

To give a formal description of the game, we assume that  $I = \{120, 80, 50\}$ . Relations  $\sim_A$  and  $\sim_B$  are the reflexive and symmetric closures of the rela-

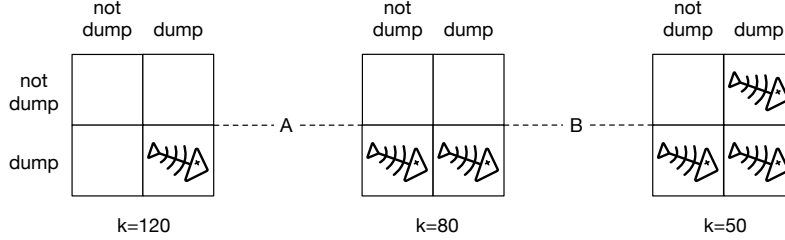


Figure 4: Game with imperfect information. In each of the three initial states, Company A selects a row and Company B selects a column.

tions  $\{(120, 80)\}$  and  $\{(80, 50)\}$ , respectively. For each initial state  $\alpha \in \{120, 80, 50\}$  and each agent  $a \in \{\text{Company A, Company B}\}$ , the set of actions  $\Delta_a^\alpha$  is the set  $\{\text{not dump, dump}\}$ . The set  $\Omega$  consists of outcomes “dead” and “alive”. The set of the plays  $P$  is defined by the content of the cells in tables in Figure 4. For example,  $(120, (\text{dump, dump}), \text{dead}) \in P$  because if  $k = 120$  and both companies dump the pollutant, then the fish die.

### 2.3. “Can do”, “doing”, and “this would do”

There are three distinct types of ability modalities that have been considered in the literature. Coalition logic [38, 39] and ATL [40] discuss modalities that represent an existential quantifier over possible actions. In such systems,  $\Box_a \varphi$  means that the agent  $a$  has an ability of achieving condition  $\varphi$  by taking a specific action. This approach does not necessarily imply that the agent will take this action. We refer to this type of modality as a “can do” modality.

The other type of modality has been considered in STIT logic where  $\Box \varphi$  refers to the *current action* of the agent [9, 10, 11, 12, 13]. Under this approach, the current action of each agent is specified as a part of the semantics. We refer to this type of modality as a “doing” modality.

Finally, there is the third approach when the action  $d$  of an agent  $a$  is specified in the syntax of the modality [41, 34, 42, 43]. In this case, formula  $\Box_{a:d} \varphi$  expresses the ability of agent  $a$  to achieve condition  $\varphi$  if she takes action  $d$ . Just like in the case of the first approach, it is not assumed that the agent will take the action  $d$ . We refer to this type of modality as a “this (action  $d$ ) would do” modality.

In the current article, we study responsibility in an imperfect information setting. All three types of modalities above could be defined in such a setting

and all three of them can be combined with the ex ante knowledge modality  $\mathbf{K}$ . If  $\Box_a\varphi$  is “can do” modality, then  $\mathbf{K}_a\Box_a\varphi$  means that agent  $a$  knows that she has an action to achieve  $\varphi$ . Note that knowing that you have an action that would guarantee a condition is different from knowing exactly which action guarantees the condition. To express the latter, one needs to introduce an epistemic version of “can do” modality, often called the “know how” modality [44, 33, 45, 30, 31, 46].

In the case of the “doing” modality, expression  $\mathbf{K}_a\Box_a\varphi$  means that agent  $a$  knows ex ante that the future action will guarantee  $\varphi$ . Since at the ex ante moment the agent does not know yet what her action will be,  $\mathbf{K}_a\Box_a\varphi$  means that the agent knows ex ante that  $\varphi$  is unavoidable. One can introduce an epistemic version of “doing” modality – “knowingly doing” [35, 37, 26]. Knowingly doing is also known as the interim knowledge modality [22]. It cannot be expressed as a combination of “doing” and ex ante knowledge modalities. We discuss the interim knowledge modality in Section 4.1.

Finally, in the case of “this would do” modality, the expression  $\mathbf{K}_a\Box_{a:d}\varphi$  states that agent  $a$  knows ex ante that action  $d$  guarantees  $\varphi$ . This expression represents the epistemic version of “this would do” modality. Different from the previous two types of abilities, it appears that there isn’t an epistemic version of “this would do” modality that is not expressible through a combination of “this would do” and the ex ante knowledge.

In a perfect information setting, the counterfactual responsibility refers to “could have done” – the “can do” ability in the past. In the same setting, the seeing-to-it responsibility is defined in terms of using an action that makes an outcome unavoidable. In this paper, we model seeing-to-it as a “doing” ability. In the imperfect information case, to define these two forms of responsibility, we use the “know how” ability and the “knowingly doing” ability. In the literature, the seeing-to-it responsibility is sometimes also defined via “this would do” ability [34, 42].

#### 2.4. Syntax and semantics

By  $\Phi^{\text{ST,CF}}$  we denote the language defined by the grammar

$$\varphi := p \mid \neg\varphi \mid \varphi \rightarrow \varphi \mid \mathbf{K}_a\varphi \mid \mathbf{ST}_a\varphi \mid \mathbf{CF}_a\varphi,$$

where  $p$  is a propositional variable and  $a \in \mathcal{A}$  is an agent. We read  $\mathbf{K}_a\varphi$  as “agent  $a$  knows ex ante that statement  $\varphi$  is true”,  $\mathbf{ST}_a\varphi$  as “agent  $a$  sees to  $\varphi$ ”, and  $\mathbf{CF}_a\varphi$  as “agent  $a$  is counterfactually responsible for  $\varphi$ ”. By  $\Phi^{\text{CF}}$  we

denote the fragment of the language  $\Phi^{\text{ST},\text{CF}}$  that does not include modality **ST**. Similarly, by  $\Phi^{\text{ST}}$  we denote the fragment of  $\Phi^{\text{ST},\text{CF}}$  without modality **CF**.

We assume that conjunction  $\wedge$  and disjunction  $\vee$ , truth  $\top$ , and false  $\perp$  are defined in the standard way. For any finite set of formulae  $X$ , by  $\wedge X$  and  $\vee X$  we mean, respectively, the disjunction and the conjunction of all formulae in set  $X$ . By definition,  $\wedge \emptyset$  and  $\vee \emptyset$  are  $\top$  and  $\perp$  respectively. Let  $\bar{K}_a$  mean  $\neg K_a \neg$ . We read  $\bar{K}_a \varphi$  as “agent  $a$  did not exclude ex ante a possibility of  $\varphi$ ”.

**Definition 2.** *The satisfaction relation  $(\alpha, \delta, \omega) \Vdash \varphi$  between a play  $(\alpha, \delta, \omega) \in P$  and a formula  $\varphi \in \Phi^{\text{ST},\text{CF}}$  is defined as:*

1.  $(\alpha, \delta, \omega) \Vdash p$ , if  $(\alpha, \delta, \omega) \in \pi(p)$ ,
2.  $(\alpha, \delta, \omega) \Vdash \neg \varphi$ , if  $(\alpha, \delta, \omega) \not\Vdash \varphi$ ,
3.  $(\alpha, \delta, \omega) \Vdash \varphi \rightarrow \psi$ , if  $(\alpha, \delta, \omega) \not\Vdash \varphi$  or  $(\alpha, \delta, \omega) \Vdash \psi$ ,
4.  $(\alpha, \delta, \omega) \Vdash K_a \varphi$ , if  $(\alpha', \delta', \omega') \Vdash \varphi$  for each play  $(\alpha', \delta', \omega') \in P$  such that  $\alpha \sim_a \alpha'$ ,
5.  $(\alpha, \delta, \omega) \Vdash \text{ST}_a \varphi$ , if
  - (a)  $(\alpha', \delta', \omega') \Vdash \varphi$  for each play  $(\alpha', \delta', \omega') \in P$  such that  $\alpha \sim_a \alpha'$  and  $\delta(a) = \delta'(a)$ ,
  - (b)  $(\alpha', \delta', \omega') \not\Vdash \varphi$  for some play  $(\alpha', \delta', \omega') \in P$  such that  $\alpha \sim_a \alpha'$ .
6.  $(\alpha, \delta, \omega) \Vdash \text{CF}_a \varphi$  if
  - (a)  $(\alpha, \delta, \omega) \Vdash \varphi$  and
  - (b) there is an action  $d \in \Delta_a^\alpha$  such that for each play  $(\alpha', \delta', \omega') \in P$  if  $\alpha \sim_a \alpha'$  and  $d = \delta'(a)$ , then  $(\alpha', \delta', \omega') \not\Vdash \varphi$ .

We illustrate Definition 2 with several examples based on the pollution game with imperfect information depicted in Figure 4.

**Example 4.**  $(80, (\text{dump}, \text{dump}), \text{dead}) \not\Vdash \text{ST}_{\text{Company } A}(\text{“Fish are dead”})$ . Indeed, consider the play  $(120, (\text{dump}, \text{not dump}), \text{alive})$ . Note that Company  $A$  cannot distinguish states 80 and 120 and also that this company uses the same action under profiles  $(\text{dump}, \text{dump})$  and  $(\text{dump}, \text{not dump})$ . Finally, observe that the fish are alive under the play  $(120, (\text{dump}, \text{not dump}), \text{alive})$ . Thus, Company  $A$  is not seeing to the death of the fish under the play  $(80, (\text{dump}, \text{dump}), \text{dead})$  by item 5(a) of Definition 2.

**Example 5.**  $(80, (dump, dump), dead) \Vdash \mathbf{CF}_{Company A}(\text{“Fish are dead”})$ . Indeed, under the profile  $(80, (dump, dump), dead)$ , the fish die and Company A had a strategy to prevent the death (“not dump”) that works not only in the current state ( $k=80$ ) but also in the other state ( $k=120$ ) that Company A cannot distinguish from the current state.

The next two examples refer to a different pollution game, which is depicted in Figure 5. In the new game, the labels A and B on the dashed lines are swapped.

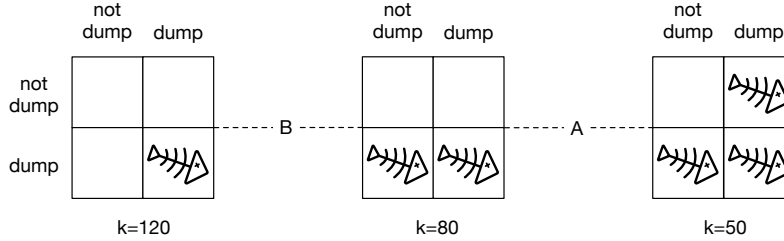


Figure 5: A modified version of Figure 4 in which the indistinguishability relations of Company A and Company B are swapped.

**Example 6.**  $(80, (dump, dump), dead) \Vdash \mathbf{ST}_{Company A}(\text{“Fish are dead”})$ . Indeed, the action (“dump”) taken by the company guarantees that the fish die not only in the current state ( $k=80$ ) but also in the other state ( $k=50$ ) that Company A cannot distinguish from the current state.

**Example 7.**  $(80, (dump, dump), dead) \not\Vdash \mathbf{CF}_{Company A}(\text{“Fish are dead”})$ . Indeed, note that although Company A had an action (“not dump”) that would have guaranteed that the fish don’t die, the company did not know this. Company A did not know because it cannot distinguish the current state ( $k=80$ ) from another state ( $k=50$ ) where the action “not dump” does not guarantee the survival of the fish.

The formal semantics in Definition 2 specifies satisfaction  $\Vdash$  as a relation between a play  $(\alpha, \delta, \omega)$  and a formula  $\varphi$ . This is different from the standard semantics of modal logics where satisfaction is a relation between a state and a formula. This change is needed because seeing-to-it (as well

as counterfactual) responsibility is a property not of a state, but rather of a play.

Indeed, note that for the game depicted in Figure 5, we have

$$\begin{aligned} (80, (\text{dump}, \text{dump}), \text{dead}) &\Vdash \mathbf{ST}_{\text{Company A}}(\text{“Fish are dead”}), \\ (120, (\text{dump}, \text{dump}), \text{dead}) &\not\Vdash \mathbf{ST}_{\text{Company A}}(\text{“Fish are dead”}). \end{aligned}$$

Thus, the existence of responsibility for seeing-to-it depends on the initial state. Also, for the same game,

$$\begin{aligned} (50, (\text{dump}, \text{dump}), \text{dead}) &\Vdash \mathbf{ST}_{\text{Company A}}(\text{“Fish are dead”}), \\ (50, (\text{not dump}, \text{dump}), \text{dead}) &\not\Vdash \mathbf{ST}_{\text{Company A}}(\text{“Fish are dead”}). \end{aligned}$$

Hence, the existence of responsibility for seeing-to-it depends on the actions. Finally, for the game depicted in Figure 1 (left), under action profile  $(m_2, d_3)$ , the mother is counterfactually responsible for the baby crying only if the baby is actually crying. So, the counterfactual responsibility depends on the outcome. For these reasons, Definition 2 specifies the satisfaction as a relation between a play (consisting of an initial state, an action profile, and an outcome) and a formula.

Next, let us turn to item 5(b) of Definition 2. It is intended to avoid agent  $a$  being responsible for unavoidable statements like  $2 + 2 = 4$ . In the perfect information case, this condition was introduced in Deliberative STIT [11]. There are at least three possible ways in which this condition can be stated in the imperfect information case:

1. agent  $a$  does not know that  $\varphi$  is unavoidable,
2.  $\varphi$  is avoidable,
3. agent  $a$  knows that  $\varphi$  is avoidable.

For example, shooting a terminally ill person by an agent who does not know that the person is about to die is seeing to the death under alternative 1, but not under alternatives 2 and 3. Out of the above alternatives, 1 is the weakest and 3 is the strongest. Item 5(b) of Definition 2 formally captures alternative 1. We believe that this treatment of responsibility is consistent with the common approach of defining legal responsibility as a combination of “guilty actions” and “guilty mind”.



### 3. Definability and Undefinability Results

This article contains three main technical results: the definability of modality **CF** through modality **ST**, the undefinability of modality **ST** through modalities **CF** and **K**, and a sound and complete axiomatisation of modality **ST**. In this section, we present the first two results. The axiomatisation is discussed in the next section.

#### 3.1. Forbearing, Refraining, and Definability of **CF** through **ST**

In this subsection, we show that the counterfactual responsibility modality **CF** is expressible through the seeing-to-it modality **ST**. In the next subsection, we prove that the opposite is false. The particular way we express modality **CF** through modality **ST** goes back to von Wright’s notion of “forbearing” [47, p.45]:

An agent, on a given occasion, forbears the doing of a certain thing if, and only if, he *can do* this thing, but *does in fact not do* it.

Belnap and Perloff suggested using the term “refraining” instead of “forbearing” [48]. They also captured the statement “agent  $a$  refrains from doing  $\varphi$ ” in STIT logic as  $\mathbf{IK}_a \neg \mathbf{IK}_a \varphi$ , where **IK** is a (non-deliberative) version of modality **ST** that does not include the *negative* condition 5(b) in Definition 2. We further discuss this modality in Section 4.1. Note that without the negative condition, formula  $\mathbf{IK}_a \neg \mathbf{IK}_a \varphi$  means that agent  $a$  acted not to see to  $\varphi$ , but it does not mean that the agent had a way to see to  $\varphi$ . However, with the negative condition, formula  $\mathbf{ST}_a \neg \mathbf{ST}_a \varphi$  implies that the agent  $a$  indeed had a way to see to  $\varphi$ . Moreover, Horty and Belnap observed that  $\mathbf{ST}_a \neg \mathbf{ST}_a \varphi$  is equivalent to the statement that agent  $a$  did not see to  $\varphi$  but had an ability to do so [11]. Although they made this observation for the perfect information case, it remains valid in the imperfect information setting of the current article. This allows us to express modality **CF** through modality **ST** as stated in the next theorem.

**Theorem 1.**  $(\alpha, \delta, \omega) \Vdash \mathbf{CF}_a \varphi$  iff  $(\alpha, \delta, \omega) \Vdash \varphi \wedge \mathbf{ST}_a \neg \mathbf{ST}_a \neg \varphi$ , for any formula  $\varphi \in \Phi^{\mathbf{ST}, \mathbf{CF}}$  and any play  $(\alpha, \delta, \omega)$  of any game.

PROOF.  $(\Rightarrow)$  : By item 6 of Definition 2, the assumption  $(\alpha, \delta, \omega) \Vdash \mathbf{CF}_a \varphi$  implies that

$$(\alpha, \delta, \omega) \Vdash \varphi \tag{1}$$

and that there is an action  $d \in \Delta_a^\alpha$  such that for each play  $(\alpha', \delta', \omega') \in P$ ,

$$\alpha \sim_a \alpha' \wedge d = \delta'(a) \Rightarrow (\alpha', \delta', \omega') \not\models \varphi. \quad (2)$$

Due to statement (1), to prove  $(\alpha, \delta, \omega) \Vdash \varphi \wedge \mathbf{ST}_a \neg \mathbf{ST}_a \neg \varphi$ , it suffices to show that

$$(\alpha, \delta, \omega) \Vdash \mathbf{ST}_a \neg \mathbf{ST}_a \neg \varphi. \quad (3)$$

We prove this statement by verifying the two claims below.

**Claim 1.**  $(\alpha'', \delta'', \omega'') \Vdash \neg \mathbf{ST}_a \neg \varphi$  for any  $(\alpha'', \delta'', \omega'') \in P$  such that  $\alpha \sim_a \alpha''$  and  $\delta(a) = \delta''(a)$ .

PROOF OF CLAIM. Statement (1) implies  $(\alpha'', \delta'', \omega'') \not\models \mathbf{ST}_a \neg \varphi$  by item 5 of Definition 2 because  $\alpha \sim_a \alpha''$  and  $\delta(a) = \delta''(a)$ . Thus,  $(\alpha'', \delta'', \omega'') \Vdash \neg \mathbf{ST}_a \neg \varphi$  by item 2 of Definition 2.  $\square$

Let  $\delta_0$  be any action profile in state  $\alpha$  such that  $\delta_0(a) = d$ . Such an action profile exists because the domain of actions of each agent in state  $\alpha$  is not empty by item 3 of Definition 1. Then, by condition 5(b) of Definition 1, there must exist at least one outcome  $\omega_0$  such that  $(\alpha, \delta_0, \omega_0) \in P$ .

**Claim 2.**  $(\alpha, \delta_0, \omega_0) \Vdash \mathbf{ST}_a \neg \varphi$ .

PROOF OF CLAIM. Recall that  $\delta_0(a) = d$  by the choice of action profile  $\delta_0$ . Thus,  $(\alpha', \delta', \omega') \not\models \varphi$  for each play  $(\alpha', \delta', \omega') \in P$  such that  $\alpha \sim_a \alpha'$  and  $\delta_0(a) = \delta'(a)$  by statement (2). At the same time,  $(\alpha, \delta, \omega) \Vdash \varphi$  by statement (1). Therefore,  $(\alpha, \delta_0, \omega_0) \Vdash \mathbf{ST}_a \neg \varphi$  by item 5 of Definition 2.  $\square$

Claims 1 and 2 imply statement (3) by item 5 of Definition 2.

( $\Leftarrow$ ): The assumption  $(\alpha, \delta, \omega) \Vdash \varphi \wedge \mathbf{ST}_a \neg \mathbf{ST}_a \neg \varphi$  implies that

$$(\alpha, \delta, \omega) \Vdash \varphi \quad (4)$$

and  $(\alpha, \delta, \omega) \Vdash \mathbf{ST}_a \neg \mathbf{ST}_a \neg \varphi$ . The latter, by item 5(b) of Definition 2, implies that there is a play  $(\alpha', \delta', \omega') \in P$  such that

$$\alpha \sim_a \alpha' \quad (5)$$

and  $(\alpha', \delta', \omega') \not\models \neg \mathbf{ST}_a \neg \varphi$ . Thus,  $(\alpha', \delta', \omega') \Vdash \mathbf{ST}_a \neg \varphi$  by item 2 of Definition 2. Then, for each play  $(\alpha'', \delta'', \omega'') \in P$ ,

$$\alpha' \sim_a \alpha'' \wedge \delta'(a) = \delta''(a) \Rightarrow (\alpha'', \delta'', \omega'') \Vdash \neg \varphi \quad (6)$$

by item 5(a) of Definition 2. Let action  $d$  be  $\delta'(a) \in \Delta_a^{\alpha'}$ . Thus,  $d \in \Delta_a^\alpha$  by item 3 of Definition 1 and statement (5). Then, by statement (5) and statement (6), for any play  $(\alpha'', \delta'', \omega'') \in P$ ,

$$\alpha \sim_a \alpha'' \wedge d = \delta''(a) \Rightarrow (\alpha'', \delta'', \omega'') \Vdash \neg\varphi.$$

Therefore,  $(\alpha, \delta, \omega) \Vdash \mathbf{CF}_a\varphi$  by item 6 of Definition 2 and statement (4).  $\square$

### 3.2. Undefinability of **ST** through **CF** and **K**

As we have seen in Theorem 1, the counterfactual responsibility modality **CF** could be defined through the seeing-to-it modality **ST**. In this section, we show that modality **ST** cannot be defined in language  $\Phi^{\mathbf{CF}}$ . To prove this, we construct two games and define a common play for both games such that a formula  $\varphi \in \Phi^{\mathbf{CF}}$  is satisfied under this play in the first game if and only if it is satisfied under the same play in the second game. We also give a formula in the language  $\Phi^{\mathbf{ST}}$  which is satisfied in the first game but not in the second game under the constructed play. The first result, in a more general form, is stated in this section as Lemma 4 and the second as Lemma 5 and Lemma 6. The undefinability is formally stated as Theorem 2 at the end of this section.

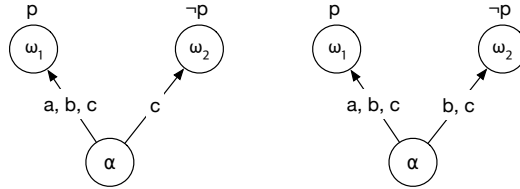


Figure 6: Two Games.

The two games are depicted in Figure 6. Each game has a single agent, Alice. In other words,  $\mathcal{A} = \{Alice\}$ . Both games have a single initial state  $\alpha$  and two outcomes:  $\omega_1$  and  $\omega_2$ . In both games, Alice has three actions ( $a$ ,  $b$ , and  $c$ ) in state  $\alpha$ . In both games, propositional variable<sup>3</sup>  $p$  holds true only in the plays that result in outcome  $\omega_1$ . In both games, action  $a$  leads to outcome  $\omega_1$  and action  $c$  leads (nondeterministically) to outcome  $\omega_1$  or

<sup>3</sup>We assume here that the language contains only one propositional variable. If the language contains more variables, the satisfaction relation of all of them should be defined the same way as  $p$ .

outcome  $\omega_2$ . The only difference between the two games is how action  $b$  is executed. In the first game, action  $b$  acts the same way as action  $a$  and in the second game it acts the same way as action  $c$ , see Figure 6.

We refer to the two games from Figure 6 as the left and the right games. The sets of plays of these two games are denoted by  $P_l$  and  $P_r$ , respectively. Satisfiability relations corresponding to those games are denoted by  $\Vdash_l$  and  $\Vdash_r$ . Valuation functions for the two games will be denoted by  $\pi_l$  and  $\pi_r$ . Note that  $\pi_l(p) = \pi_r(p) = \{(\alpha, x, \omega_1) \mid x \in \{a, b, c\}\}$  by the choice of the games.

Recall that an action profile is a function that maps agents into actions. Since Alice is the only agent in these two games, we refer to an action profile by the action of Alice under the profile. The play  $(\alpha, b, \omega_1)$  is the common play of these two games that we use to show the undefinability of modality **ST** in language  $\Phi^{\text{CF}}$ .

As mentioned above, Lemma 4 is a key step in the proof of undefinability. Before proving this lemma, we establish three auxiliary results. First, observe that sets  $P_l$  and  $P_r$  are not equal because  $(\alpha, b, \omega_2) \in P_r \setminus P_l$ . However, the set of plays that use actions  $a$  and  $c$  is the same for both games.

**Lemma 1.**  $(\alpha, \delta, \omega) \in P_l$  iff  $(\alpha, \delta, \omega) \in P_r$  for any action  $\delta \in \{a, c\}$  and any outcome  $\omega \in \{\omega_1, \omega_2\}$ .  $\square$

Next, note that plays  $(\alpha, a, \omega_1)$  and  $(\alpha, b, \omega_1)$  in the left game are indistinguishable in language  $\Phi^{\text{CF}}$ :

**Lemma 2.**  $(\alpha, a, \omega_1) \Vdash_l \varphi$  iff  $(\alpha, b, \omega_1) \Vdash_l \varphi$  for each formula  $\varphi \in \Phi^{\text{CF}}$ .  $\square$

**PROOF.** We prove the lemma by structural induction on formula  $\varphi$ . First, suppose that  $\varphi$  is propositional variable  $p$ . Note that  $(\alpha, a, \omega_1) \in \pi_l(p)$  and  $(\alpha, b, \omega_1) \in \pi_l(p)$  by the choice of valuation function  $\pi_l$ . Thus,  $(\alpha, a, \omega_1) \Vdash_l p$  and  $(\alpha, b, \omega_1) \Vdash_l p$  by item 1 of Definition 2.

If formula  $\varphi$  is a negation or an implication, then the required follows from the induction hypothesis and items 2 and 3 of Definition 2.

Next, suppose that formula  $\varphi$  has the form  $\mathbf{K}_{\text{Alice}}\psi$ . Without loss of generality, assume that  $(\alpha, a, \omega_1) \Vdash_l \mathbf{K}_{\text{Alice}}\psi$ . Then,  $(\alpha, \delta, \omega) \Vdash_l \psi$  for each play  $(\alpha, \delta, \omega) \in P_l$  by item 4 of Definition 2. Therefore,  $(\alpha, b, \omega_1) \Vdash_l \mathbf{K}_{\text{Alice}}\psi$  again by item 4 of Definition 2 and because there is only one state that Alice cannot distinguish from state  $\alpha$  – the state  $\alpha$  itself.

Finally, assume that formula  $\varphi$  has the form  $\mathbf{CF}_{\text{Alice}}\psi$ . Without loss of generality, assume that  $(\alpha, a, \omega_1) \Vdash_l \mathbf{CF}_{\text{Alice}}\psi$ . Thus, by item 6 of Definition 2,

1.  $(\alpha, a, \omega_1) \Vdash_l \psi$  and
2. there is an action  $x \in \{a, b, c\}$  such that  $(\alpha, x, \omega) \not\Vdash_l \psi$  for each play  $(\alpha, x, \omega) \in P_l$ .

Item 1 above implies  $(\alpha, b, \omega_1) \Vdash_l \psi$  by the induction hypothesis. Therefore,  $(\alpha, b, \omega_1) \Vdash_l \mathbf{CF}_{Alice}\psi$  by item 2 above and the same item 6 of Definition 2.  $\boxtimes$

Similarly, the actions  $b$  and  $c$  in the right game are indistinguishable in language  $\Phi^{\mathbf{CF}}$ :

**Lemma 3.**  $(\alpha, b, \omega) \Vdash_r \varphi$  iff  $(\alpha, c, \omega) \Vdash_r \varphi$  for each outcome  $\omega \in \{\omega_1, \omega_2\}$  and each formula  $\varphi \in \Phi^{\mathbf{CF}}$ .  $\boxtimes$

The next lemma is one of the two key steps in the proof of undefinability. It shows that for any common play of the two games, the same formulae are satisfied in this play under both games. Of course, play  $(\alpha, b, \omega_2) \in P_r \setminus P_l$  is excluded.

**Lemma 4.**  $(\alpha, \delta, \omega) \Vdash_l \varphi$  iff  $(\alpha, \delta, \omega) \Vdash_r \varphi$  for each formula  $\varphi \in \Phi^{\mathbf{CF}}$  and each play  $(\alpha, \delta, \omega) \in P_l$ .

**PROOF.** We prove the lemma by structural induction on formula  $\varphi$ . If  $\varphi$  is a propositional variable  $p$ , then  $(\alpha, \delta, \omega) \Vdash_l p$  iff  $(\alpha, \delta, \omega) \Vdash_r p$  by Definition 2 and because  $\pi_l(p) = \pi_r(p)$ . The case when formula  $\varphi$  is a negation or an implication follows from the induction hypothesis and items 2 and 3 of Definition 2 respectively.

Suppose that formula  $\varphi$  has the form  $\mathbf{K}_{Alice}\psi$ . Recall that there is only one state that Alice cannot distinguish from state  $\alpha$  – the state  $\alpha$  itself.

$(\Rightarrow)$  : Assume that  $(\alpha, \delta, \omega) \Vdash_l \mathbf{K}_{Alice}\psi$ . Thus,  $(\alpha, \delta', \omega') \Vdash_l \psi$  for each play  $(\alpha, \delta', \omega') \in P_l$  by item 4 of Definition 2. Hence, by the induction hypothesis,  $(\alpha, \delta', \omega') \Vdash_r \psi$  for each play  $(\alpha, \delta', \omega') \in P_l$ . Then,  $(\alpha, \delta', \omega') \Vdash_r \psi$  for each play  $(\alpha, \delta', \omega') \in P_r$  because  $P_r = P_l \cup \{(\alpha, b, \omega_2)\}$  and Lemma 3 entails that  $(\alpha, c, \omega_2) \Vdash_r \psi$  implies  $(\alpha, b, \omega_2) \Vdash_r \psi$ .

$(\Leftarrow)$  : Suppose that  $(\alpha, \delta, \omega) \Vdash_r \mathbf{K}_{Alice}\psi$ . Then,  $(\alpha, \delta', \omega') \Vdash_r \psi$  for each play  $(\alpha, \delta', \omega') \in P_r$  by item 4 of Definition 2. Hence,  $(\alpha, \delta', \omega') \Vdash_r \psi$  for each play  $(\alpha, \delta', \omega') \in P_l$  because  $P_l \subseteq P_r$ . Thus, by the induction hypothesis,  $(\alpha, \delta', \omega') \Vdash_l \psi$  for each play  $(\alpha, \delta', \omega') \in P_l$ . Therefore,  $(\alpha, \delta, \omega) \Vdash_l \mathbf{K}_{Alice}\psi$  by item 4 of Definition 2.

Finally, suppose that formula  $\varphi$  has the form  $\mathbf{CF}_{Alice}\psi$ .

$(\Rightarrow)$  : Let  $(\alpha, \delta, \omega) \Vdash_l \mathbf{CF}_{Alice}\psi$  for some play  $(\alpha, \delta, \omega) \in P_l$ . Hence, by item 6 of Definition 2,

$$(\alpha, \delta, \omega) \Vdash_l \psi \quad (7)$$

and

$$\exists x \in \{a, b, c\} \forall (\alpha, x, \omega') \in P_l ((\alpha, x, \omega') \not\Vdash_l \psi).$$

Thus, by Lemma 2,

$$\exists x \in \{a, c\} \forall (\alpha, x, \omega') \in P_l ((\alpha, x, \omega') \not\Vdash_l \psi).$$

Then, by the induction hypothesis,

$$\exists x \in \{a, c\} \forall (\alpha, x, \omega') \in P_l ((\alpha, x, \omega') \not\Vdash_r \psi).$$

Hence, by Lemma 1,

$$\exists x \in \{a, c\} \forall (\alpha, x, \omega') \in P_r ((\alpha, x, \omega') \not\Vdash_r \psi).$$

In addition,  $(\alpha, \delta, \omega) \Vdash_r \psi$  also by the induction hypothesis using statement (7). Therefore,  $(\alpha, \delta, \omega) \Vdash_r \mathbf{CF}_{Alice}\psi$  by item 6 of Definition 2.

$(\Leftarrow)$  : Suppose  $(\alpha, \delta, \omega) \Vdash_r \mathbf{CF}_{Alice}\psi$ . Thus, by item 6 of Definition 2,

$$(\alpha, \delta, \omega) \Vdash_r \psi \quad (8)$$

and

$$\exists x \in \{a, b, c\} \forall (\alpha, x, \omega') \in P_r ((\alpha, x, \omega') \not\Vdash_r \psi).$$

Then, by Lemma 3,

$$\exists x \in \{a, c\} \forall (\alpha, x, \omega') \in P_r ((\alpha, x, \omega') \not\Vdash_r \psi).$$

Hence, by Lemma 1,

$$\exists x \in \{a, c\} \forall (\alpha, x, \omega') \in P_l ((\alpha, x, \omega') \not\Vdash_r \psi).$$

Thus, by the induction hypothesis,

$$\exists x \in \{a, c\} \forall (\alpha, x, \omega') \in P_l ((\alpha, x, \omega') \not\Vdash_l \psi).$$

In addition,  $(\alpha, \delta, \omega) \Vdash_l \psi$  also by the induction hypothesis using statement (8). Therefore,  $(\alpha, \delta, \omega) \Vdash_l \mathbf{CF}_{Alice}\psi$  by item 6 of Definition 2.  $\square$

Informally, the next two lemmas are true because by choosing action  $b$  in the left model the agent knows that statement  $p$  will be unavoidably true, while the same is not true about the right model. Formally, statements of the lemmas follow from the definitions of the left and the right models and item 5 of Definition 2.

**Lemma 5.**  $(\alpha, b, \omega_1) \Vdash_l \mathbf{ST}_{Alice} p.$  ☒

**Lemma 6.**  $(\alpha, b, \omega_1) \not\K_r \mathbf{ST}_{Alice} p.$  ☒

The statement of the next theorem follows from Lemma 4, Lemma 5, and Lemma 6.

**Theorem 2.** *Modality  $\mathbf{ST}$  is not definable in the language  $\Phi^{\mathbf{CF}}$ .*

#### 4. Logical System

In this section, we present our third main result – a sound and complete logical system for the seeing-to-it modality.

##### 4.1. Two Forms of Knowledge

It is common in game theory to distinguish *ex ante*, *interim*, and *ex post* knowledge [22]. The first refers to the knowledge of an agent before an action is taken, the second to the knowledge at the moment of taking an action, and the third to the knowledge after an action is taken. The knowledge modality  $\mathbf{K}$ , as defined in item 4 of Definition 2, captures *ex ante* (before action) knowledge. The *interim* knowledge (at the moment of action) can be captured by modality  $\mathbf{IK}$  defined by removing condition (b) from item 5 of Definition 2:

**Definition 3.**  $(\alpha, \delta, \omega) \Vdash \mathbf{IK}_a \varphi$ , if  $(\alpha', \delta', \omega') \Vdash \varphi$  for each play  $(\alpha', \delta', \omega') \in P$  such that  $\alpha \sim_a \alpha'$  and  $\delta(a) = \delta'(a)$ ,

In STIT logic, modality  $\mathbf{IK}$  is usually referred to as “non-deliberative seeing-to-it” in contrast to “deliberative seeing-to-it” captured by modality  $\mathbf{ST}$ . Modality  $\mathbf{IK}$  has been studied in perfect information [9, 10, 11] and imperfect information settings [35, 37, 26, 12]. It has all the standard S5 properties plus the Independence of Agency property captured by the following inference rule:

$$\frac{\neg\varphi_1 \vee \dots \vee \neg\varphi_n}{\neg\mathbf{IK}_{a_1}\varphi_1 \vee \dots \vee \neg\mathbf{IK}_{a_n}\varphi_n},$$

where agents  $a_1, \dots, a_n$  are pairwise different. Informally, this rule says that if statements  $\varphi_1, \dots, \varphi_n$  are logically inconsistent, then, in any given state, at least one of the agents  $a_i$  cannot see to the corresponding condition  $\varphi_i$ .

Note that each of our responsibility modalities can be defined through a combination of the ex ante and the interim knowledge modalities:

$$\begin{aligned}\mathbf{ST}_a\varphi &= \mathbf{IK}_a\varphi \wedge \overline{\mathbf{K}}_a\neg\varphi, \\ \mathbf{CF}_a\varphi &= \varphi \wedge \overline{\mathbf{K}}_a\mathbf{IK}_a\neg\varphi.\end{aligned}$$

As mentioned above, various axiomatisations of modality  $\mathbf{IK}$  have been proposed before [24, 25, 26]. To explicitly capture the properties of the seeing-to-it responsibility, in this section, we give an axiomatisation of the modality  $\mathbf{ST}$  without decomposing it into two knowledge modalities. In addition to  $\mathbf{ST}$ , our logical system also includes the ex ante knowledge modality  $\mathbf{K}$ . We do not include the modality  $\mathbf{CF}$  because, as shown in Theorem 1, it is definable through modality  $\mathbf{ST}$ .

#### 4.2. Axioms and Inference Rules

In the rest of this section, we present a sound and complete logical system that describes the interplay between modalities  $\mathbf{K}$  and  $\mathbf{ST}$  in language  $\Phi^{\mathbf{ST}}$ .

In addition to tautologies in language  $\Phi^{\mathbf{ST}}$ , our logical system contains the following axioms:

1. Truth:  $\mathbf{K}_a\varphi \rightarrow \varphi$  and  $\mathbf{ST}_a\varphi \rightarrow \varphi$ ,
2. Negative Introspection:  $\neg\mathbf{K}_a\varphi \rightarrow \mathbf{K}_a\neg\mathbf{K}_a\varphi$ ,
3. Distributivity:  $\mathbf{K}_a(\varphi \rightarrow \psi) \rightarrow (\mathbf{K}_a\varphi \rightarrow \mathbf{K}_a\psi)$ ,
4. Introspection of Responsibility:  $\mathbf{ST}_a\varphi \rightarrow \mathbf{ST}_a\mathbf{ST}_a\varphi$ ,
5. Refraining:  $\neg\mathbf{ST}_a\varphi \wedge \overline{\mathbf{K}}_a\mathbf{ST}_a\varphi \rightarrow \mathbf{ST}_a\neg\mathbf{ST}_a\varphi$ ,
6. Dual Responsibility:  $\mathbf{ST}_a\varphi \wedge \mathbf{ST}_a\psi \rightarrow \mathbf{ST}_a(\varphi \wedge \psi)$ ,
7. Non-Responsibility for Known:  $\mathbf{K}_a\varphi \rightarrow \neg\mathbf{ST}_a\varphi$ .

We write  $\vdash \varphi$  if formulae  $\varphi$  is derivable in our logical system using the following rules of inference: Indirect Responsibility

$$\frac{\mathbf{K}_a\psi \wedge \bigwedge_{i \leq n} \mathbf{ST}_a\chi_i \rightarrow \varphi}{\mathbf{K}_a\psi \wedge \bigwedge_{i \leq n} \mathbf{ST}_a\chi_i \rightarrow \mathbf{K}_a\varphi \vee \mathbf{ST}_a\varphi},$$

the Independence of Agency (for distinct agents  $b_1, \dots, b_n$ )

$$\frac{\bigwedge_{i \leq m} \mathbf{K}_{a_i}\varphi_i \rightarrow \bigvee_{j \leq n} \neg\psi_j}{\bigwedge_{i \leq m} \mathbf{K}_{a_i}\varphi_i \rightarrow \bigvee_{j \leq n} \mathbf{K}_{b_j}\neg\mathbf{ST}_{b_j}\psi_j},$$



the Modus Ponens, and the Necessitation

$$\frac{\varphi, \varphi \rightarrow \psi}{\psi} \qquad \frac{\varphi}{\mathbf{K}_a\varphi}.$$

We say that  $\varphi$  is a *theorem* of our logical system if  $\vdash \varphi$ .

In addition to the unary relation  $\vdash \varphi$ , we also consider binary relation  $X \vdash \varphi$  between a set of formulae  $X$  and a formula  $\varphi$ . Let  $X \vdash \varphi$  if formula  $\varphi$  is derivable from the *theorems* of our logical system and the set of additional axioms  $X$  using *only* the Modus Ponens inference rule. It is easy to see that statement  $\emptyset \vdash \varphi$  is equivalent to  $\vdash \varphi$ . We say that set  $X$  is consistent if  $X \not\vdash \perp$ .

**Lemma 7 (Lindenbaum).** *Any consistent set of formulae in the language  $\Phi^{\text{ST}}$  can be extended to a maximal consistent set of formulae.*

PROOF. The standard proof of Lindenbaum's lemma [49, Proposition 2.14] applies here too.  $\square$

The next four lemmas state well-known properties of S5 modality  $\mathbf{K}$ . Their proofs can be found in the appendix.

**Lemma 8 (Deduction).** *If  $X, \varphi \vdash \psi$ , then  $X \vdash \varphi \rightarrow \psi$ .*

**Lemma 9 (Positive Introspection).**  $\vdash \mathbf{K}_a\varphi \rightarrow \mathbf{K}_a\mathbf{K}_a\varphi$ .

**Lemma 10.** *If  $\varphi_1, \dots, \varphi_n \vdash \psi$ , then  $\mathbf{K}_a\varphi_1, \dots, \mathbf{K}_a\varphi_n \vdash \mathbf{K}_a\psi$ .*

**Lemma 11.**  $\vdash \mathbf{K}_a\varphi_1 \wedge \dots \wedge \mathbf{K}_a\varphi_n \leftrightarrow \mathbf{K}_a(\varphi_1 \wedge \dots \wedge \varphi_n)$ , if  $n \geq 0$ .

### 4.3. Soundness

In this subsection, we prove the soundness of our logical system. The soundness of the Truth axiom for modality  $\mathbf{K}$ , the Negative Introspection axiom, the Distributivity axiom, the Modus Ponens inference rule, and the Necessitation inference rule is straightforward. We show the soundness of the remaining axioms and inference rules as separate lemmas. We start by observing two auxiliary results that follow from items 4 and 5 of Definition 2.

**Lemma 12.** *For any plays  $(\alpha, \delta, \omega)$  and  $(\alpha', \delta', \omega')$  of an arbitrary game, if  $(\alpha, \delta, \omega) \Vdash \mathbf{K}_a\varphi$  and  $\alpha \sim_a \alpha'$ , then  $(\alpha', \delta', \omega') \Vdash \mathbf{K}_a\varphi$ .*  $\square$

**Lemma 13.** *For any plays  $(\alpha, \delta, \omega)$  and  $(\alpha', \delta', \omega')$  of an arbitrary game, if  $(\alpha, \delta, \omega) \Vdash \mathbf{ST}_a\varphi$ ,  $\alpha \sim_a \alpha'$ , and  $\delta(a) = \delta'(a)$ , then  $(\alpha', \delta', \omega') \Vdash \mathbf{ST}_a\varphi$ .  $\boxtimes$*

The Truth axiom for modality  $\mathbf{ST}$  states that if an agent sees to  $\varphi$ , then statement  $\varphi$  is true.

**Lemma 14.** *If  $(\alpha, \delta, \omega) \Vdash \mathbf{ST}_a\varphi$ , then  $(\alpha, \delta, \omega) \Vdash \varphi$ .*

PROOF. By item 5(a) of Definition 2, the assumption  $(\alpha, \delta, \omega) \Vdash \mathbf{ST}_a\varphi$  implies that  $(\alpha', \delta', \omega') \Vdash \varphi$  for each play  $(\alpha', \delta', \omega') \in P$  such that  $\alpha \sim_a \alpha'$  and  $\delta(a) = \delta'(a)$ . In particular,  $(\alpha, \delta, \omega) \Vdash \varphi$ .  $\boxtimes$

The Introspection of Responsibility axiom states that if an agent sees to  $\varphi$ , then she sees to it that she sees to  $\varphi$ . The statement of this axiom is not as straightforward as it might appear due to the epistemic negative condition 5(b) in Definition 2.

**Lemma 15.** *If  $(\alpha, \delta, \omega) \Vdash \mathbf{ST}_a\varphi$ , then  $(\alpha, \delta, \omega) \Vdash \mathbf{ST}_a\mathbf{ST}_a\varphi$ .*

PROOF. To prove  $(\alpha, \delta, \omega) \Vdash \mathbf{ST}_a\mathbf{ST}_a\varphi$ , we verify conditions (a) and (b) of item 5 in Definition 2.

*Condition a:* Consider any play  $(\alpha', \delta', \omega') \in P$  such that  $\alpha \sim_a \alpha'$  and  $\delta(a) = \delta'(a)$ . It suffices to show that  $(\alpha', \delta', \omega') \Vdash \mathbf{ST}_a\varphi$ , which, by Lemma 13, follows from the assumption  $(\alpha, \delta, \omega) \Vdash \mathbf{ST}_a\varphi$  of the current lemma.

*Condition b:* By item 5(b) of Definition 2, the assumption  $(\alpha, \delta, \omega) \Vdash \mathbf{ST}_a\varphi$  of the lemma implies that  $(\alpha', \delta', \omega') \not\Vdash \varphi$  for some play  $(\alpha', \delta', \omega') \in P$  such that  $\alpha \sim_a \alpha'$ . Hence,  $(\alpha', \delta', \omega') \not\Vdash \mathbf{ST}_a\varphi$  by Lemma 14.  $\boxtimes$

The Refraining axiom states that if an agent does not see to  $\varphi$  and she did not exclude ex ante a possibility of seeing to  $\varphi$ , then she refrains from (forbears) doing  $\varphi$ . See Section 3.1 for the discussion of refraining/forbearing.

**Lemma 16.** *If  $(\alpha, \delta, \omega) \not\Vdash \mathbf{ST}_a\varphi$  and  $(\alpha, \delta, \omega) \Vdash \overline{\mathbf{K}}_a\mathbf{ST}_a\varphi$ , then  $(\alpha, \delta, \omega) \Vdash \mathbf{ST}_a\neg\mathbf{ST}_a\varphi$ .*

PROOF. By the definition of modality  $\overline{\mathbf{K}}$  and items 2 and 4 of Definition 2, the assumption  $(\alpha, \delta, \omega) \Vdash \overline{\mathbf{K}}_a\mathbf{ST}_a\varphi$  implies that there is a play  $(\alpha_1, \delta_1, \omega_1) \in P$  such that

$$\alpha \sim_a \alpha_1 \tag{9}$$

and

$$(\alpha_1, \delta_1, \omega_1) \Vdash \mathbf{ST}_a \varphi. \quad (10)$$

By item 5(b) of Definition 2, statement (10) implies that

$$\exists (\alpha', \delta', \omega') \in P (\alpha_1 \sim_a \alpha' \wedge (\alpha', \delta', \omega') \not\Vdash \varphi). \quad (11)$$

By item 5 of Definition 2, assumption  $(\alpha, \delta, \omega) \not\Vdash \mathbf{ST}_a \varphi$  of the lemma implies that *at least one* of the following conditions holds:

1. there is a play  $(\alpha', \delta', \omega') \in P$  such that  $\alpha \sim_a \alpha'$ ,  $\delta(a) = \delta'(a)$ , and  $(\alpha', \delta', \omega') \not\Vdash \varphi$ ,
2.  $(\alpha', \delta', \omega') \Vdash \varphi$  for each play  $(\alpha', \delta', \omega') \in P$  such that  $\alpha \sim_a \alpha'$ .

Note that the second of the above statements is not consistent with statements (9) and (11). Hence, the first of the above statements is true. In other words, there is a play  $(\alpha_2, \delta_2, \omega_2) \in P$  such that

$$\alpha \sim_a \alpha_2 \wedge \delta(a) = \delta_2(a) \wedge (\alpha_2, \delta_2, \omega_2) \not\Vdash \varphi. \quad (12)$$

**Claim 3.**  $(\alpha', \delta', \omega') \Vdash \neg \mathbf{ST}_a \varphi$  for each play  $(\alpha', \delta', \omega') \in P$  such that  $\alpha \sim_a \alpha'$  and  $\delta(a) = \delta'(a)$ .

**PROOF OF CLAIM.** By item 2 of Definition 2, it suffices to show that  $(\alpha', \delta', \omega') \not\Vdash \mathbf{ST}_a \varphi$ . To prove this, by item 5(a) of Definition 2, it is enough to establish that  $\alpha' \sim_a \alpha_2$ ,  $\delta'(a) = \delta_2(a)$ , and  $(\alpha_2, \delta_2, \omega_2) \not\Vdash \varphi$ . All three of these statements follow from statement (12) and the assumptions  $\alpha \sim_a \alpha'$  and  $\delta(a) = \delta'(a)$  of the claim.  $\square$

Claim 3, statement (9), and statement (10) imply  $(\alpha, \delta, \omega) \Vdash \mathbf{ST}_a \neg \mathbf{ST}_a \varphi$  by item 5 of Definition 2.  $\square$

The Dual Responsibility axiom states that if an agent sees to both  $\varphi$  and  $\psi$ , then she sees to the conjunction  $\varphi \wedge \psi$ .

**Lemma 17.** *If  $(\alpha, \delta, \omega) \Vdash \mathbf{ST}_a \varphi$  and  $(\alpha, \delta, \omega) \Vdash \mathbf{ST}_a \psi$ , then  $(\alpha, \delta, \omega) \Vdash \mathbf{ST}_a(\varphi \wedge \psi)$ .*

**PROOF.** The assumption  $(\alpha, \delta, \omega) \Vdash \mathbf{ST}_a \varphi$ , by item 5 of Definition 2, implies

$$\forall (\alpha', \delta', \omega') \in P (\alpha \sim_a \alpha' \wedge \delta(a) = \delta'(a) \Rightarrow ((\alpha', \delta', \omega') \Vdash \varphi)) \quad (13)$$

and

$$\exists(\alpha', \delta', \omega') \in P (\alpha \sim_a \alpha' \wedge ((\alpha', \delta', \omega') \not\models \varphi)). \quad (14)$$

Similarly, the assumption  $(\alpha, \delta, \omega) \Vdash \mathbf{ST}_a \psi$ , by item 5(a) of the same Definition 2, implies

$$\forall(\alpha', \delta', \omega') \in P (\alpha \sim_a \alpha' \wedge \delta(a) = \delta'(a) \Rightarrow ((\alpha', \delta', \omega') \Vdash \psi)). \quad (15)$$

By Definition 2 and the definition of connective  $\wedge$ , statements (13) and (15) imply

$$\forall(\alpha', \delta', \omega') \in P (\alpha \sim_a \alpha' \wedge \delta(a) = \delta'(a) \Rightarrow ((\alpha', \delta', \omega') \Vdash \varphi \wedge \psi)). \quad (16)$$

Similarly, by Definition 2 and the definition of connective  $\wedge$ , statement (14) implies

$$\exists(\alpha', \delta', \omega') \in P (\alpha \sim_a \alpha' \wedge ((\alpha', \delta', \omega') \not\models \varphi \wedge \psi)). \quad (17)$$

Finally, note that  $(\alpha, \delta, \omega) \Vdash \mathbf{ST}_a(\varphi \wedge \psi)$  follows from statement (16) and statement (17) by item 5 of Definition 2.  $\square$

The Non-Responsibility for Known axiom states that an agent does not see to something about which she knows *ex ante* that it is guaranteed to happen.

**Lemma 18.** *If  $(\alpha, \delta, \omega) \Vdash \mathbf{K}_a \varphi$ , then  $(\alpha, \delta, \omega) \not\models \mathbf{ST}_a \varphi$ .*

PROOF. By item 4 of Definition 2, assumption  $(\alpha, \delta, \omega) \Vdash \mathbf{K}_a \varphi$  implies that  $(\alpha', \delta', \omega') \Vdash \varphi$  for each play  $(\alpha', \delta', \omega') \in P$  such that  $\alpha \sim_a \alpha'$ . Therefore,  $(\alpha, \delta, \omega) \not\models \mathbf{ST}_a \varphi$  by item 5(b) of Definition 2.  $\square$

Next, assume that a statement  $\varphi$  holds true each time when an agent  $a$  is responsible for statements  $\chi_1, \dots, \chi_n$ . Then, in each such situation, agent  $a$  is also responsible for  $\varphi$  unless the agent  $a$  knows *ex ante* that  $\varphi$  is unavoidably true. This is captured by the Indirect Responsibility inference rule. Note that the actual Indirect Responsibility rule is slightly more general because it includes an additional assumption  $\mathbf{K}_a \psi$ .

**Lemma 19.** *If formula  $\mathbf{K}_a \psi \wedge \bigwedge_{i \leq n} \mathbf{ST}_a \chi_i \rightarrow \varphi$  is satisfied on each play of each game, then formula*

$$\mathbf{K}_a \psi \wedge \bigwedge_{i \leq n} \mathbf{ST}_a \chi_i \rightarrow \mathbf{K}_a \varphi \vee \mathbf{ST}_a \varphi$$

*is also satisfied on each play of each game.*

PROOF. Suppose that there is a play  $(\alpha, \delta, \omega) \in P$  of a game

$$(I, \{\sim_a\}_{a \in \mathcal{A}}, \{\Delta_a^\alpha\}_{a \in \mathcal{A}}^{\alpha \in I}, \Omega, P, \pi)$$

such that

$$(\alpha, \delta, \omega) \Vdash \mathbf{K}_a \psi, \quad (18)$$

$$(\alpha, \delta, \omega) \Vdash \mathbf{ST}_a \chi_i, \quad \forall i \leq n, \quad (19)$$

$$(\alpha, \delta, \omega) \not\Vdash \mathbf{K}_a \varphi, \quad (20)$$

$$(\alpha, \delta, \omega) \not\Vdash \mathbf{ST}_a \varphi. \quad (21)$$

By item 4 of Definition 2, statement (20) implies that there exists a play  $(\alpha_1, \delta_1, \omega_1) \in P$  such that

$$(\alpha_1, \delta_1, \omega_1) \not\Vdash \varphi \text{ and } \alpha \sim_a \alpha_1. \quad (22)$$

At the same time, by item 5 of Definition 2, statement (21) implies that *at least one* of the following is true:

1.  $(\alpha', \delta', \omega') \not\Vdash \varphi$  for some  $(\alpha', \delta', \omega') \in P$  such that  $\alpha \sim_a \alpha'$  and  $\delta(a) = \delta'(a)$ ,
2.  $(\alpha', \delta', \omega') \Vdash \varphi$  for each  $(\alpha', \delta', \omega') \in P$  such that  $\alpha \sim_a \alpha'$ .

The second of the above conditions is not consistent with statement (22). Thus, the first statement must be true. In other words, there is a play  $(\alpha_2, \delta_2, \omega_2) \in P$  such that  $\alpha \sim_a \alpha_2$ ,  $\delta(a) = \delta_2(a)$ , and

$$(\alpha_2, \delta_2, \omega_2) \not\Vdash \varphi. \quad (23)$$

Thus, by Lemma 12 and Lemma 13, statements (18) and (19) imply

$$\begin{aligned} (\alpha_2, \delta_2, \omega_2) \Vdash \mathbf{K}_a \psi, \\ (\alpha_2, \delta_2, \omega_2) \Vdash \mathbf{ST}_a \chi_i, \quad \forall i \leq n. \end{aligned}$$

Therefore,  $(\alpha_2, \delta_2, \omega_2) \Vdash \varphi$  by the assumption of the lemma, which contradicts statement (23).  $\boxtimes$

Informally, the Independence of Agency rule is true because, by item 5 of Definition 1, for each action profile there is at least one possible outcome. See the proof below for more details.

**Lemma 20.** *If  $\bigwedge_{i \leq m} \mathbf{K}_{a_i} \varphi_i \rightarrow \bigvee_{j \leq n} \neg \psi_j$  is satisfied on each play of each game, then  $\bigwedge_{i \leq m} \mathbf{K}_{a_i} \varphi_i \rightarrow \bigvee_{j \leq n} \mathbf{K}_{b_j} \neg \mathbf{ST}_{b_j} \psi_j$  is also satisfied on each play of each game, where agents  $b_1, \dots, b_n$  are distinct.*

PROOF. Suppose that there is a play  $(\alpha, \delta, \omega)$  of a game such that  $(\alpha, \delta, \omega) \Vdash \bigwedge_{i \leq m} \mathbf{K}_{a_i} \varphi_i$  and  $(\alpha, \delta, \omega) \Vdash \neg \mathbf{K}_{b_j} \neg \mathbf{ST}_{b_j} \psi_j$  for each  $j \leq n$ . The latter, by items 2 and 4 of Definition 2, implies that for each  $j \leq n$  there is a play  $(\alpha'_j, \delta'_j, \omega'_j)$  such that  $\alpha \sim_{b_j} \alpha'_j$  and

$$(\alpha'_j, \delta'_j, \omega'_j) \Vdash \mathbf{ST}_{b_j} \psi_j. \quad (24)$$

By Definition 1, set  $\Delta_a^\alpha$  contains at least one element  $d_a$ . Consider the action profile

$$\widehat{\delta}(a) = \begin{cases} \delta'_j(b_j), & \text{if } a = b_j, \\ d_a, & \text{otherwise.} \end{cases} \quad (25)$$

Such a profile is well-defined because agents  $b_1, \dots, b_n$  are distinct. By item 5 of Definition 1, there must exist an outcome  $\widehat{\omega}$  such that  $(\alpha, \widehat{\delta}, \widehat{\omega})$  is a play. Note that, by Lemma 12, the assumption  $(\alpha, \delta, \omega) \Vdash \bigwedge_{i \leq m} \mathbf{K}_{a_i} \varphi_i$  implies that  $(\alpha, \widehat{\delta}, \widehat{\omega}) \Vdash \bigwedge_{i \leq m} \mathbf{K}_{a_i} \varphi_i$ . Then,  $(\alpha, \widehat{\delta}, \widehat{\omega}) \Vdash \bigvee_{j \leq n} \neg \psi_j$  by the assumption of the lemma. Thus, there must exist  $j_0 \leq n$  such that

$$(\alpha, \widehat{\delta}, \widehat{\omega}) \Vdash \neg \psi_{j_0}. \quad (26)$$

At the same time,  $\alpha'_{j_0} \sim_{b_{j_0}} \alpha$  by the choice of play  $(\alpha'_{j_0}, \delta'_{j_0}, \omega'_{j_0})$ . Also,  $\delta'_{j_0}(b_{j_0}) = \widehat{\delta}(b_{j_0})$  by equation (25). Thus,  $(\alpha, \widehat{\delta}, \widehat{\omega}) \Vdash \psi_{j_0}$ , by item 5(a) of Definition 2 and statement (24), which contradicts statement (26).  $\square$

The lemmas above imply the following strong soundness theorem.

**Theorem 3 (Strong Soundness).** *For any play  $(\alpha, \delta, \omega)$  of any game, if  $(\alpha, \delta, \omega) \Vdash \chi$  for each formula  $\chi \in X$  and  $X \vdash \varphi$ , then  $(\alpha, \delta, \omega) \Vdash \varphi$ .*

#### 4.4. Towards the Proof of Completeness

In this subsection, we give the intuition behind the proof of completeness of our logical system. The actual proof is in the next subsection.

As usual, the proof of completeness consists of a construction of a canonical model. In our case, it is a canonical game. The key component of our proof of completeness, just like most other such proofs, is the “induction” (or “truth”) lemma – Lemma 26 in this article. In the case of many classical

modal logics, such as S5, the induction lemma states that a formula is satisfied in a state if and only if it belongs to the maximal consistent set that defines this state:  $w \Vdash \varphi$  iff  $\varphi \in w$ . In our case, the induction lemma states that a formula is satisfied by a play if and only if it belongs to the outcome of the play:  $(\alpha, \delta, \omega) \Vdash \varphi$  iff  $\varphi \in \omega$ . Thus, all properties of a play expressible in language  $\Phi^{\text{ST}}$  are determined by the outcome alone and not by the initial state  $\alpha$  or the action profile  $\delta$ . In other words, the outcomes of the canonical game “remember” all the relevant information about the initial state and the action profile. This is a peculiar property of our canonical game. It is not true for the class of all games, as specified in Definition 1. However, the fact that the canonical game belongs to a subclass of games from Definition 1 makes our completeness result stronger, not weaker.

There are several ways in which the completeness theorem for our logical system differs from the proof of completeness for the epistemic logic of counterfactual responsibility [20], where it is called *blameworthiness*. The most significant difference is how the set of actions  $\Delta_a^\alpha$  is defined. In the case of counterfactual responsibility, for each play  $(\alpha, \delta, \omega)$  and each formula  $\mathbf{CF}_a\varphi \in \omega$ , the induction lemma in [20] requires that  $(\alpha, \delta, \omega) \Vdash \mathbf{CF}_a\varphi$ . To achieve this, the canonical model is equipped with an action that, when executed by an agent  $a$ , prevents  $\varphi$ . Such an action is called  $\neg\varphi$  and the set of plays of the game is defined so that by “voting for” action  $\neg\varphi$ , agent  $a$  is able to prevent formula  $\varphi$  from being true. Informally, in a canonical game from [20], each agent votes for a formula which is meant to be satisfied in the outcome.

The situation is quite different in the current setting. Indeed, consider any two plays  $(\alpha, \delta, \omega)$  and  $(\alpha', \delta', \omega')$  such that  $\alpha \sim_a \alpha'$  and  $\delta(a) = \delta'(a)$ . Then, by Lemma 13, for any formula  $\varphi \in \Phi^{\text{ST}}$ ,

$$(\alpha, \delta, \omega) \Vdash \mathbf{ST}_a\varphi \quad \text{iff} \quad (\alpha', \delta', \omega') \Vdash \mathbf{ST}_a\varphi.$$

In other words, agent  $a$  sees to the same statements in both plays. This means that in each initial state  $\alpha'$  indistinguishable to agent  $a$  from initial state  $\alpha$ , if the agent acts the same way as she acts under  $\delta(a)$ , then she sees to *everything* that she sees to in play  $(\alpha, \delta, \omega)$ . Informally, by acting the same way, agent  $a$  says that she accepts the responsibility for everything she sees to in play  $(\alpha, \delta, \omega)$ . Recall from our earlier discussion that, in our canonical game, the validity of all formulae, including the formulae of the form  $\mathbf{ST}_a\varphi$ , is completely determined by the outcome  $\omega$ . Thus, by acting the same way

as in play  $(\alpha, \delta, \omega)$ , agent  $a$  says “make me responsible for everything I am responsible in outcome  $\omega$ ”. To capture this intuition, instead of making the action (“vote”) of agent  $a$  to be a particular formula, we use the whole set  $\omega$  as the action.

Now consider an arbitrary action profile  $\delta$  that assigns each agent  $a$  an action (outcome)  $\delta(a)$  for which she votes. The game aggregates votes of all agents into a single outcome  $\omega$  in which each agent  $a$  is responsible for exactly the same formulae as those in outcome  $\delta(a)$ . This aggregation mechanism is specified in Definition 6, which defines the set of all plays of the canonical game. Although the definition only requires that if  $\mathbf{ST}_a\varphi \in \delta(a)$ , then  $\mathbf{ST}_a\varphi \in \omega$ , it can be shown that the converse is also true in the canonical model. In a similar fashion, the initial states are also defined as outcomes. Informally, if the game starts in an initial state  $\omega$ , then it means that outcome  $\omega$  is potentially reachable from this initial state. Recall that knowledge modality  $\mathbf{K}$  refers to the *ex ante* knowledge or knowledge in the initial state. Thus, all outcomes that can be reached from the same initial state must have the same  $\mathbf{K}$ -formulae. We formally capture this in item 1 of Definition 6. Although this item only requires that if  $\mathbf{K}_a\varphi \in \alpha$ , then  $\mathbf{K}_a\varphi \in \omega$ , it again can be shown that the converse is true in the canonical model.

#### 4.5. Completeness

In this subsection, we prove the completeness of our logical system following the outline from the previous subsection. We start by defining the canonical game  $(\Omega, \{\sim_a\}_{a \in \mathcal{A}}, \{\Delta_a^\alpha\}_{a \in \mathcal{A}}, \Omega, P, \pi)$ , where  $\Omega$  is the set of all maximal consistent sets of formulae.

Since modality  $\mathbf{K}$  captures *ex ante* knowledge, we define two initial states to be indistinguishable by an agent  $a$  if the sets have the same  $\mathbf{K}_a$ -formulae.

**Definition 4.** *For any two initial states  $\alpha, \alpha' \in \Omega$  and any agent  $a \in \mathcal{A}$ , let  $\alpha \sim_a \alpha'$  when  $\mathbf{K}_a\varphi \in \alpha$  iff  $\mathbf{K}_a\varphi \in \alpha'$  for each formula  $\varphi \in \Phi^{\mathbf{ST}}$ .*

As we discussed in the previous subsection, actions of an agent  $a$  are outcomes which have the same  $\mathbf{K}_a$ -formulae as the initial state  $\alpha$ . Taking Definition 4 into account, this can be stated as follows:

**Definition 5.** *Let  $\Delta_a^\alpha = \{\omega \in \Omega \mid \alpha \sim_a \omega\}$ , for any initial state  $\alpha \in \Omega$  and any agent  $a \in \mathcal{A}$ .*



Note that Definition 4 and Definition 5 imply that set  $\Delta_a^\alpha$  satisfies the uniformity condition of Definition 1.

Recall that the outcome inherits the *ex ante* knowledge from the initial state and the responsibility from the actions of the individual agents:

**Definition 6.** *The set of plays  $P$  contains all triples  $(\alpha, \delta, \omega)$ , where  $\alpha \in \Omega$  is an initial state,  $\delta$  is an action profile in state  $\alpha$ , and  $\omega \in \Omega$  is an outcome, such that for each agent  $a \in \mathcal{A}$  and each formula  $\varphi \in \Phi^{\text{ST}}$ ,*

1. *if  $\mathbf{K}_a\varphi \in \alpha$ , then  $\mathbf{K}_a\varphi \in \omega$ ,*
2. *if  $\mathbf{ST}_a\varphi \in \delta(a)$ , then  $\mathbf{ST}_a\varphi \in \omega$ .*

Finally, as we discussed earlier, the validity of all formulae in our canonical game, including propositional variables, is completely determined by the outcome:

**Definition 7.**  $\pi(p) = \{(\alpha, \delta, \omega) \in P \mid p \in \omega\}$ .

This concludes the definition of the canonical model. The next lemma verifies the condition from item 5 of Definition 1.

**Lemma 21.** *For any initial state  $\alpha \in \Omega$  and any action profile  $\delta$  in state  $\alpha$ , there is an outcome  $\omega \in \Omega$  such that  $(\alpha, \delta, \omega) \in P$ .*

PROOF. Consider the set of formulae

$$X = \{\mathbf{K}_a\varphi \mid \mathbf{K}_a\varphi \in \alpha, a \in \mathcal{A}\} \cup \{\mathbf{ST}_a\psi \mid \mathbf{ST}_a\psi \in \delta(a), a \in \mathcal{A}\}.$$

First, we show that this set is consistent. Suppose the opposite. Then, there are agents  $a_1, \dots, a_n$  and formulae

$$\mathbf{K}_{a_1}\varphi_1, \dots, \mathbf{K}_{a_n}\varphi_n \in \alpha \tag{27}$$

as well as *distinct* agents  $b_1, \dots, b_m$  and formulae

$$\begin{aligned} &\mathbf{ST}_{b_1}\psi_1^1, \dots, \mathbf{ST}_{b_1}\psi_1^{k_1} \in \delta(b_1) \\ &\dots \\ &\mathbf{ST}_{b_m}\psi_m^1, \dots, \mathbf{ST}_{b_m}\psi_m^{k_m} \in \delta(b_m) \end{aligned} \tag{28}$$

such that

$$k_1, \dots, k_m \geq 1 \tag{29}$$

and

$$\mathbf{K}_{a_1}\varphi_1, \dots, \mathbf{K}_{a_n}\varphi_n, \mathbf{ST}_{b_1}\psi_1^1, \dots, \mathbf{ST}_{b_1}\psi_1^{k_1}, \dots, \mathbf{ST}_{b_m}\psi_m^1, \dots, \mathbf{ST}_{b_m}\psi_m^{k_m} \vdash \perp.$$

Thus, by Lemma 8 and propositional reasoning,

$$\vdash \left( \bigwedge_{i \leq n} \mathbf{K}_{a_i}\varphi_i \right) \wedge \left( \bigwedge_{i \leq m} \bigwedge_{j \leq k_i} \mathbf{ST}_{b_i}\psi_i^j \right) \rightarrow \perp.$$

Then, again by propositional reasoning,

$$\vdash \bigwedge_{i \leq n} \mathbf{K}_{a_i}\varphi_i \rightarrow \neg \bigwedge_{i \leq m} \bigwedge_{j \leq k_i} \mathbf{ST}_{b_i}\psi_i^j.$$

Thus, by De Morgan's law,

$$\vdash \bigwedge_{i \leq n} \mathbf{K}_{a_i}\varphi_i \rightarrow \bigvee_{i \leq m} \neg \bigwedge_{j \leq k_i} \mathbf{ST}_{b_i}\psi_i^j.$$

Hence, by the Independence of Agency inference rule,

$$\vdash \bigwedge_{i \leq n} \mathbf{K}_{a_i}\varphi_i \rightarrow \bigvee_{i \leq m} \mathbf{K}_{b_i} \neg \mathbf{ST}_{b_i} \bigwedge_{j \leq k_i} \mathbf{ST}_{b_i}\psi_i^j.$$

Thus, by propositional reasoning using statement (27),

$$\alpha \vdash \bigvee_{i \leq m} \mathbf{K}_{b_i} \neg \mathbf{ST}_{b_i} \bigwedge_{j \leq k_i} \mathbf{ST}_{b_i}\psi_i^j.$$

Then, because set  $\alpha$  is maximal, there exists  $i_0 \leq m$  such that

$$\mathbf{K}_{b_{i_0}} \neg \mathbf{ST}_{b_{i_0}} \bigwedge_{j \leq k_{i_0}} \mathbf{ST}_{b_{i_0}}\psi_{i_0}^j \in \alpha.$$

The assumption of the lemma that  $\delta$  is an action profile in state  $\alpha$  implies  $\delta(b_{i_0}) \in \Delta_{b_{i_0}}^\alpha$  by item 5(a) of Definition 1. Hence,  $\alpha \sim_{b_{i_0}} \delta(b_{i_0})$  by Definition 5. Thus, by Definition 4,

$$\mathbf{K}_{b_{i_0}} \neg \mathbf{ST}_{b_{i_0}} \bigwedge_{j \leq k_{i_0}} \mathbf{ST}_{b_{i_0}}\psi_{i_0}^j \in \delta(b_{i_0}).$$

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

$$\delta(b_{i_0}) \vdash \neg \mathbf{ST}_{b_{i_0}} \bigwedge_{j \leq k_{i_0}} \mathbf{ST}_{b_{i_0}} \psi_{i_0}^j. \quad (30)$$

At the same time, by the Introspection of Responsibility axiom and the Modus Ponens inference rule, assumptions (28) imply

$$\begin{aligned} \delta(b_{i_0}) \vdash \mathbf{ST}_{b_{i_0}} \mathbf{ST}_{b_{i_0}} \psi_{i_0}^1, \\ \dots \\ \delta(b_{i_0}) \vdash \mathbf{ST}_{b_{i_0}} \mathbf{ST}_{b_{i_0}} \psi_{i_0}^{k_{i_0}}. \end{aligned}$$

Thus, by propositional reasoning using the Dual Responsibility axiom and assumption (29),

$$\delta(b_{i_0}) \vdash \mathbf{ST}_{b_{i_0}} \bigwedge_{j \leq k_{i_0}} \mathbf{ST}_{b_{i_0}} \psi_{i_0}^j,$$

which contradicts statement (30) because set  $\delta_\alpha(b_{i_0})$  is consistent. Therefore, set  $X$  is consistent.

Let  $\omega$  be any maximal consistent extension of set  $X$ . Such a set  $\omega$  exists by Lemma 7. Then,  $(\alpha, \delta, \omega) \in P$  by Definition 6 and the choice of set  $X$ .  $\square$

A key step in a proof of the completeness theorem is usually an ‘‘induction’’ or ‘‘truth’’ lemma. In our case, it is Lemma 26. As we mentioned in the previous subsection, this lemma states that  $(\alpha, \delta, \omega) \Vdash \psi$  iff  $\psi \in \omega$ . The next four lemmas are auxiliary lemmas used to prove the induction steps of this lemma for formula  $\psi$  of different forms. The first of them is used to prove ( $\Leftarrow$ ) direction in the case when the formula has the form  $\mathbf{ST}_a \varphi$ .

**Lemma 22.** *If  $(\alpha, \delta, \omega) \in P$  and  $\mathbf{ST}_a \varphi \in \omega$ , then*

1. *for each play  $(\alpha', \delta', \omega') \in P$ , if  $\alpha \sim_a \alpha'$  and  $\delta(a) = \delta'(a)$ , then  $\varphi \in \omega'$ , and*
2. *there is  $(\alpha', \delta', \omega') \in P$  such that  $\alpha \sim_a \alpha'$  and  $\varphi \notin \omega'$ .*

**PROOF.** We prove statements 1 and 2 separately.

**Statement 1.** Consider any play  $(\alpha', \delta', \omega') \in P$  such that  $\alpha \sim_a \alpha'$  and

$$\delta(a) = \delta'(a). \quad (31)$$

Suppose that  $\varphi \notin \omega'$ . Then,  $\omega' \not\vdash \varphi$  because set  $\omega'$  is maximal. Hence,  $\mathbf{ST}_a\varphi \notin \omega'$  by the contraposition of the Truth axiom. Then,  $\mathbf{ST}_a\varphi \notin \delta'(a)$  by Definition 6 and the assumption  $(\alpha', \delta', \omega') \in P$ . Thus,  $\mathbf{ST}_a\varphi \notin \delta(a)$  because of equation (31). Hence,

$$\neg\mathbf{ST}_a\varphi \in \delta(a) \quad (32)$$

because set  $\delta(a)$  is maximal. At the same time, formula  $\mathbf{K}_a\neg\mathbf{ST}_a\varphi \rightarrow \neg\mathbf{ST}_a\varphi$  is an instance of the Truth axiom. Hence,  $\vdash \mathbf{ST}_a\varphi \rightarrow \neg\mathbf{K}_a\neg\mathbf{ST}_a\varphi$  by contraposition. Then,  $\omega \vdash \neg\mathbf{K}_a\neg\mathbf{ST}_a\varphi$  by the assumption  $\mathbf{ST}_a\varphi \in \omega$  of the lemma. Thus,  $\mathbf{K}_a\neg\mathbf{ST}_a\varphi \notin \omega$  because set  $\omega$  is consistent. Hence, by Definition 6 and assumption  $(\alpha, \delta, \omega) \in P$  of the lemma,

$$\mathbf{K}_a\neg\mathbf{ST}_a\varphi \notin \alpha. \quad (33)$$

Note that  $\delta(a) \in \Delta_a^\alpha$  by item 5(a) of Definition 1. Thus,  $\alpha \sim_a \delta(a)$  by Definition 5. Hence,  $\mathbf{K}_a\neg\mathbf{ST}_a\varphi \notin \delta(a)$  by Definition 4 and assumption (33). Then,  $\neg\mathbf{K}_a\neg\mathbf{ST}_a\varphi \in \delta(a)$  because set  $\delta(a)$  is maximal. Then,  $\bar{\mathbf{K}}_a\mathbf{ST}_a\varphi \in \delta(a)$  by the definition of modality  $\bar{\mathbf{K}}$ . Thus,  $\delta(a) \vdash \mathbf{ST}_a\neg\mathbf{ST}_a\varphi$  by the Refraining axiom using statement (32) and propositional reasoning. Then,  $\mathbf{ST}_a\neg\mathbf{ST}_a\varphi \in \delta(a)$  because set  $\delta(a)$  is maximal. Hence,  $\mathbf{ST}_a\neg\mathbf{ST}_a\varphi \in \omega$  by Definition 6 and the assumption  $(\alpha, \delta, \omega) \in P$  of the lemma. Then,  $\omega \vdash \neg\mathbf{ST}_a\varphi$  by the Truth axiom and the Modus Ponens inference rule. Therefore,  $\mathbf{ST}_a\varphi \notin \omega$  because set  $\omega$  is consistent, which contradicts the assumption  $\mathbf{ST}_a\varphi \in \omega$  of the lemma.

**Statement 2.** Let  $X = \{\mathbf{K}_a\psi \mid \mathbf{K}_a\psi \in \alpha\} \cup \{\neg\varphi\}$ . First, we show that set  $X$  is consistent. Assume the opposite. Then, there are formulae  $\mathbf{K}_a\psi_1, \dots, \mathbf{K}_a\psi_n \in \alpha$  such that  $\mathbf{K}_a\psi_1, \dots, \mathbf{K}_a\psi_n \vdash \varphi$ . Hence,  $\mathbf{K}_a\mathbf{K}_a\psi_1, \dots, \mathbf{K}_a\mathbf{K}_a\psi_n \vdash \mathbf{K}_a\varphi$  by Lemma 10. Then,  $\mathbf{K}_a\psi_1, \dots, \mathbf{K}_a\psi_n \vdash \mathbf{K}_a\varphi$  by Lemma 9 and the Modus Ponens inference rule. Thus,  $\alpha \vdash \mathbf{K}_a\varphi$  by the choice of formulae  $\mathbf{K}_a\psi_1, \dots, \mathbf{K}_a\psi_n$ . Hence,  $\mathbf{K}_a\varphi \in \alpha$  because set  $\alpha$  is maximal. Then,  $\mathbf{K}_a\varphi \in \omega$  by Definition 6 and the assumption  $(\alpha, \delta, \omega) \in P$  of the lemma. Thus,  $\omega \vdash \neg\mathbf{ST}_a\varphi$  by the Non-Responsibility for Known axiom and the Modus Ponens inference rule. Thus,  $\mathbf{ST}_a\varphi \notin \omega$  because set  $\omega$  is consistent, which contradicts assumption  $\mathbf{ST}_a\varphi \in \omega$  of the lemma. Therefore, set  $X$  is consistent.

Let set  $\omega'$  be any maximal consistent extension of set  $X$ . Such set  $\omega'$  exists by Lemma 7. Then,  $\neg\varphi \in X \subseteq \omega'$ . Thus,  $\varphi \notin \omega'$  because set  $\omega'$  is consistent. Define  $\alpha'$  to be  $\omega'$  and action profile  $\delta'$  to be such that  $\delta'(a) = \omega'$  for each agent  $a \in \mathcal{A}$ . Therefore,  $(\alpha', \delta', \omega') \in P$  by Definition 6.

**Claim 4.**  $\alpha \sim_a \alpha'$ .

PROOF OF CLAIM. By Definition 4 and the choice of  $\alpha'$ , it suffices to show that  $\mathbf{K}_a\psi \in \alpha$  iff  $\mathbf{K}_a\psi \in \omega'$  for each formula  $\psi \in \Phi^{\text{ST}}$ . First, if  $\mathbf{K}_a\psi \in \alpha$ , then  $\mathbf{K}_a\psi \in X \subseteq \omega'$  by the choice of sets  $X$  and  $\omega'$ . Second, suppose  $\mathbf{K}_a\psi \notin \alpha$ . Thus,  $\neg\mathbf{K}_a\psi \in \alpha$  because set  $\alpha$  is maximal. Then,  $\alpha \vdash \mathbf{K}_a\neg\mathbf{K}_a\psi$  by the Negative Introspection axiom and the Modus Ponens inference rule. Hence,  $\mathbf{K}_a\neg\mathbf{K}_a\psi \in \alpha$  because set  $\alpha$  is maximal. Thus,  $\mathbf{K}_a\neg\mathbf{K}_a\psi \in X \subseteq \omega'$  by the choice of sets  $X$  and  $\omega'$ . Then,  $\omega' \vdash \neg\mathbf{K}_a\psi$  by the Truth axiom and the Modus Ponens inference rule. Therefore,  $\mathbf{K}_a\psi \notin \omega'$  because set  $\omega'$  is consistent.  $\square$

This concludes the proof of the lemma.  $\square$

The next lemma is used in Lemma 26 to prove  $(\Rightarrow)$  direction in the case when a formula has the form  $\text{ST}_a\varphi$ .

**Lemma 23.** *If  $(\alpha, \delta, \omega) \in P$  and  $\text{ST}_a\varphi \notin \omega$ , then at least one of the following statements is true:*

1. *there exists a play  $(\alpha', \delta', \omega') \in P$  such that  $\alpha \sim_a \alpha'$ ,  $\delta(a) = \delta'(a)$ , and  $\varphi \notin \omega'$ ,*
2.  *$\varphi \in \omega'$  for each play  $(\alpha', \delta', \omega') \in P$  such that  $\alpha \sim_a \alpha'$ .*

PROOF. We consider the following two cases separately.

**Case I:**  $\mathbf{K}_a\varphi \notin \alpha$ . We show that statement 1 of the lemma holds. Consider the following set of formulae

$$X = \{\neg\varphi\} \cup \{\mathbf{K}_a\psi \mid \mathbf{K}_a\psi \in \alpha\} \cup \{\text{ST}_a\chi \mid \text{ST}_a\chi \in \delta(a)\}.$$

**Claim 5.** *Set  $X$  is consistent.*

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

$$\mathbf{K}_a\psi_1, \dots, \mathbf{K}_a\psi_n \in \alpha \tag{34}$$

and formulae

$$\text{ST}_a\chi_1, \dots, \text{ST}_a\chi_m \in \delta(a) \tag{35}$$

such that, using Lemma 8,

$$\vdash \bigwedge_{i \leq n} \mathbf{K}_a\psi_i \wedge \bigwedge_{j \leq m} \text{ST}_a\chi_j \rightarrow \varphi.$$

Then, by propositional reasoning using Lemma 11,

$$\vdash \mathbf{K}_a \left( \bigwedge_{i \leq n} \psi_i \right) \wedge \bigwedge_{j \leq m} \mathbf{ST}_a \chi_j \rightarrow \varphi.$$

Thus, by the Indirect Responsibility inference rule,

$$\vdash \mathbf{K}_a \left( \bigwedge_{i \leq n} \psi_i \right) \wedge \bigwedge_{j \leq m} \mathbf{ST}_a \chi_j \rightarrow \mathbf{K}_a \varphi \vee \mathbf{ST}_a \varphi.$$

Hence, by propositional reasoning using statement (35) and Lemma 11,

$$\delta(a) \vdash \bigwedge_{i \leq n} \mathbf{K}_a \psi_i \rightarrow \mathbf{K}_a \varphi \vee \mathbf{ST}_a \varphi.$$

Also, note that  $\delta(a) \in \Delta_a^\alpha$  by item 5(a) of Definition 1 and the assumption  $(\alpha, \delta, \omega) \in P$ . Then,  $\alpha \sim_a \delta(a)$  by Definition 5. Thus,  $\mathbf{K}_a \psi_1, \dots, \mathbf{K}_a \psi_n \in \delta(a)$  by Definition 4 and assumption (34). Then, by propositional reasoning,

$$\delta(a) \vdash \mathbf{K}_a \varphi \vee \mathbf{ST}_a \varphi. \quad (36)$$

Recall that  $\mathbf{K}_a \varphi \notin \alpha$  by the assumption of the case. Hence,  $\mathbf{K}_a \varphi \notin \delta(a)$  by Definition 4 and because  $\alpha \sim_a \delta(a)$ . Thus,  $\neg \mathbf{K}_a \varphi \in \delta(a)$  because set  $\delta(a)$  is maximal. Then,  $\delta(a) \vdash \mathbf{ST}_a \varphi$  from statement (36) by propositional reasoning. Hence,  $\mathbf{ST}_a \varphi \in \delta(a)$  again because set  $\delta(a)$  is maximal. Thus,  $\mathbf{ST}_a \varphi \in \omega$  by Definition 6 and the assumption  $(\alpha, \delta, \omega) \in P$  of the lemma, which contradicts the other assumption of the lemma:  $\mathbf{ST}_a \varphi \notin \omega$ .  $\square$

Let  $\omega'$  be any maximal consistent extension of set  $X$ . Such a set  $\omega'$  exists by Lemma 7. Note that  $\neg \varphi \in X \subseteq \omega'$ . Thus,  $\varphi \notin \omega'$  because set  $\omega'$  is consistent. Let  $\alpha'$  be set  $\omega'$ .

The proof of the following statement is identical to the proof of Claim 4.

**Claim 6.**  $\alpha \sim_a \alpha'$ .  $\square$

Define an action profile  $\delta'$  to be such that

$$\delta'(b) = \begin{cases} \delta(a), & \text{if } b = a, \\ \omega' = \alpha', & \text{otherwise.} \end{cases} \quad (37)$$

**Claim 7.**  $(\alpha', \delta', \omega') \in P$ .

PROOF OF CLAIM. We need to verify conditions 1 and 2 from Definition 6. To show the first condition, consider any formula  $\psi \in \Phi^{\text{ST}}$  such that  $\mathbf{K}_a\psi \in \alpha'$ . Thus,  $\mathbf{K}_a\psi \in \omega'$  because  $\alpha' = \omega'$  by the choice of  $\alpha'$ . To verify the second condition, it suffices to show that if  $\mathbf{ST}_b\chi \in \delta'(b)$ , then  $\mathbf{ST}_b\chi \in \omega'$  for each formula  $\chi \in \Phi^{\text{ST}}$  and each agent  $b \in \mathcal{A}$ . Indeed, assume  $\mathbf{ST}_b\chi \in \delta'(b)$  and consider two cases:

*Case (i):*  $b = a$ . Note that  $\delta'(a) = \delta(a)$  by equation (37). Then,  $\mathbf{ST}_a\chi \in \delta(a)$  by the assumption  $\mathbf{ST}_b\chi \in \delta'(b)$ . Therefore, by the choice of sets  $X$  and  $\omega'$ , we have  $\mathbf{ST}_a\chi \in X \subseteq \omega'$ .

*Case (ii):*  $b \neq a$ . Then,  $\delta'(b) = \omega'$  by equation (37). Therefore,  $\mathbf{ST}_b\chi \in \omega'$  by the assumption  $\mathbf{ST}_b\chi \in \delta'(b)$ .  $\square$

**Case II:**  $\mathbf{K}_a\varphi \in \alpha$ . We show that statement 2 of the lemma holds. Consider any play  $(\alpha', \delta', \omega') \in P$  such that  $\alpha \sim_a \alpha'$ . It suffices to show that  $\varphi \in \omega'$ . Indeed, assumptions  $\mathbf{K}_a\varphi \in \alpha$  and  $\alpha \sim_a \alpha'$  imply that  $\mathbf{K}_a\varphi \in \alpha'$  by Definition 4. Then,  $\mathbf{K}_a\varphi \in \omega'$  by Definition 6 and the assumption  $(\alpha', \delta', \omega') \in P$ . Thus,  $\omega' \vdash \varphi$  by the Truth axiom and the Modus Ponens inference rule. Therefore,  $\varphi \in \omega'$  because set  $\omega'$  is maximal.  $\square$

The next lemma is used in Lemma 26 to prove the ( $\Leftarrow$ ) direction in the case when a formula has the form  $\mathbf{K}_a\varphi$ .

**Lemma 24.** *If  $(\alpha, \delta, \omega) \in P$  and  $\mathbf{K}_a\varphi \in \omega$ , then  $\varphi \in \omega'$  for each  $(\alpha', \delta', \omega') \in P$  such that  $\alpha \sim_a \alpha'$ .*

PROOF. Suppose that  $\varphi \notin \omega'$ . Thus,  $\neg\varphi \in \omega'$  because set  $\omega'$  is maximal. Hence,  $\omega' \vdash \neg\mathbf{K}_a\varphi$  by the contraposition of the Truth axiom and the Modus Ponens inference rule. Then,  $\mathbf{K}_a\varphi \notin \omega'$  because set  $\omega'$  is consistent. Thus,  $\mathbf{K}_a\varphi \notin \alpha'$  by Definition 6 and the assumption  $(\alpha', \delta', \omega') \in P$ . Hence,  $\mathbf{K}_a\varphi \notin \alpha$  by Definition 4 and the assumption  $\alpha \sim_a \alpha'$ . Then,  $\neg\mathbf{K}_a\varphi \in \alpha$  because set  $\alpha$  is maximal. Thus,  $\alpha \vdash \mathbf{K}_a\neg\mathbf{K}_a\varphi$  by the Negative Introspection axiom and the Modus Ponens inference rule. Hence,  $\mathbf{K}_a\neg\mathbf{K}_a\varphi \in \alpha$  because set  $\alpha$  is maximal. Then,  $\mathbf{K}_a\neg\mathbf{K}_a\varphi \in \omega$  by Definition 6 and the assumption  $(\alpha, \delta, \omega) \in P$ . Thus,  $\omega \vdash \neg\mathbf{K}_a\varphi$  by the Truth axiom and the Modus Ponens inference rule. Therefore,  $\mathbf{K}_a\varphi \notin \omega$  because set  $\omega$  is consistent.  $\square$

The next lemma is used in Lemma 26 to prove ( $\Rightarrow$ ) direction in the case when a formula has the form  $\mathbf{K}_a\varphi$ .

**Lemma 25.** *If  $(\alpha, \delta, \omega) \in P$  and  $\mathbf{K}_a\varphi \notin \omega$ , then there is a play  $(\alpha', \delta', \omega') \in P$  such that  $\alpha \sim_a \alpha'$  and  $\varphi \notin \omega'$ .*

PROOF. Let  $X = \{\neg\varphi\} \cup \{\mathbf{K}_a\psi \mid \mathbf{K}_a\psi \in \alpha\}$ . First, we show that set  $X$  is consistent. Suppose the opposite. Then, there are formulae  $\mathbf{K}_a\psi_1, \dots, \mathbf{K}_a\psi_n \in \alpha$  where  $\mathbf{K}_a\psi_1, \dots, \mathbf{K}_a\psi_n \vdash \varphi$ . Hence,  $\mathbf{K}_a\mathbf{K}_a\psi_1, \dots, \mathbf{K}_a\mathbf{K}_a\psi_n \vdash \mathbf{K}_a\varphi$  by Lemma 10. Then,  $\mathbf{K}_a\psi_1, \dots, \mathbf{K}_a\psi_n \vdash \mathbf{K}_a\varphi$  by Lemma 9 and the Modus Ponens inference rule. Thus,  $\alpha \vdash \mathbf{K}_a\varphi$  by the choice of formulae  $\mathbf{K}_a\psi_1, \dots, \mathbf{K}_a\psi_n$ . Then,  $\mathbf{K}_a\varphi \in \alpha$  because set  $\alpha$  is maximal. Therefore,  $\mathbf{K}_a\varphi \in \omega$  by Definition 6 and the assumption  $(\alpha, \delta, \omega) \in P$  of the lemma, which contradicts the assumption  $\mathbf{K}_a\varphi \notin \omega$  of the lemma. Therefore, set  $X$  is consistent.

Let  $\alpha'$  be any maximal consistent extension of set  $X$ . Such set  $\alpha'$  exists by Lemma 7.

**Claim 8.**  $\alpha \sim_a \alpha'$ .

PROOF OF CLAIM. By Definition 4, it suffices to show that  $\mathbf{K}_a\psi \in \alpha$  iff  $\mathbf{K}_a\psi \in \alpha'$  for each formula  $\psi$ . If  $\mathbf{K}_a\psi \in \alpha$ , then  $\mathbf{K}_a\psi \in X \subseteq \alpha'$  by the choice of sets  $X$  and  $\alpha'$ .

Suppose that  $\mathbf{K}_a\psi \notin \alpha$ . We will show that  $\mathbf{K}_a\psi \notin \alpha'$ . Indeed, the assumption  $\mathbf{K}_a\psi \notin \alpha$  implies that  $\neg\mathbf{K}_a\psi \in \alpha$  because set  $\alpha$  is maximal. Hence,  $\alpha \vdash \mathbf{K}_a\neg\mathbf{K}_a\psi$  by the Negative Introspection axiom and the Modus Ponens inference rule. Then,  $\mathbf{K}_a\neg\mathbf{K}_a\psi \in \alpha$  again because set  $\alpha$  is maximal. Thus,  $\mathbf{K}_a\neg\mathbf{K}_a\psi \in X \subseteq \alpha'$  by the choice of sets  $X$  and  $\alpha'$ . Hence,  $\alpha' \vdash \neg\mathbf{K}_a\psi$  by the Truth axiom and the Modus Ponens inference rule. Therefore,  $\mathbf{K}_a\psi \notin \alpha'$  because set  $\alpha'$  is maximal.  $\square$

Consider action profile  $\delta'$  such that  $\delta'(b) = \alpha'$  for each agent  $b \in \mathcal{A}$ . Let  $\omega' = \alpha'$ . Note that  $\neg\varphi \in X \subseteq \alpha' = \omega'$ . Thus,  $\varphi \notin \omega'$  because set  $\omega'$  is consistent.

**Claim 9.**  $(\alpha', \delta', \omega') \in P$ .

PROOF OF CLAIM. We will prove the two items of Definition 6. First, suppose that  $\mathbf{K}_b\psi \in \alpha'$  for some formula  $\mathbf{K}_b\psi$ . Thus,  $\mathbf{K}_b\psi \in \omega'$  because  $\omega' = \alpha'$ .



Second, assume that  $\mathbf{ST}_b\psi \in \delta'(b)$ . It suffices to show that  $\mathbf{ST}_b\psi \in \omega'$ . The assumption  $\mathbf{ST}_b\psi \in \delta'(b)$  implies  $\mathbf{ST}_b\psi \in \alpha'$  again by the choice of action profile  $\delta'$ . Therefore,  $\mathbf{ST}_b\psi \in \omega'$  by the choice of outcome  $\omega'$ .  $\square$

This concludes the proof of the lemma.  $\square$

**Lemma 26 (Truth Lemma).**  $(\alpha, \delta, \omega) \Vdash \varphi$  iff  $\varphi \in \omega$ .

PROOF. We prove the lemma by induction on the structural complexity of formula  $\varphi$ . The base case follows from item 1 of Definition 2 and Definition 7. The cases when formula  $\varphi$  is an implication or a negation follow from the induction hypothesis, items 2 and 3 of Definition 2, and the maximality and consistency of set  $\omega$  in the standard way. If formula  $\varphi$  has the form  $\mathbf{K}_a\psi$ , then the required follows from item 4 of Definition 2, the induction hypothesis, Lemma 25, and Lemma 24. Finally, if formula  $\varphi$  has the form  $\mathbf{ST}_a\psi$ , then the required follows from item 5 of Definition 2, the induction hypothesis, Lemma 23, and Lemma 22.  $\square$

**Theorem 4 (Strong Completeness).** *If  $X \not\vdash \varphi$ , then there is a play  $(\alpha, \delta, \omega)$  of a game such that  $(\alpha, \delta, \omega) \Vdash \chi$  for each formula  $\chi \in X$  and  $(\alpha, \delta, \omega) \not\vdash \varphi$ .*

PROOF. Suppose that  $X \not\vdash \varphi$ . Consider any maximal consistent extension  $\omega$  of set  $X \cup \{\neg\varphi\}$ . Such set  $\omega$  exists by Lemma 7. Set  $\omega$  is an outcome of the canonical game  $(\Omega, \{\sim_a\}_{a \in \mathcal{A}}, \{\Delta_a^\alpha\}_{a \in \mathcal{A}}^{\alpha \in \Omega}, \Omega, P, \pi)$ . Let  $\alpha = \omega$  and  $\delta$  be such an action profile that  $\delta(a) = \omega$  for each agent  $a \in \mathcal{A}$ . Hence,  $(\alpha, \delta, \omega) \in P$  by Definition 6.

Also,  $\chi \in X \subseteq \omega$  for each formula  $\chi \in X$ . Thus,  $(\alpha, \delta, \omega) \Vdash \chi$  for each formula  $\chi \in X$  by Lemma 26. Similarly,  $(\alpha, \delta, \omega) \Vdash \neg\varphi$ . Therefore,  $(\alpha, \delta, \omega) \not\vdash \varphi$  by item 2 of Definition 2.  $\square$

## 5. Conclusion

In this article, we studied the ex ante knowledge modality and two responsibility modalities: “seeing-to-it” and “counterfactually responsible”. We observed that “counterfactually responsible” could be defined through “seeing-to-it” and have shown that the converse is not true. We also gave a

sound and complete axiomatization of the interplay between the individual knowledge and the “seeing-to-it” modalities.

The natural next step is to generalise the “seeing-to-it” modality to capture *group* responsibility. This, however, is not a trivial step as a faithful definition of “seeing-to-it as a group” must require a certain level of coordination and information sharing between the members of the group. Going back to our example depicted in the left diagram of Figure 1, if the mom and the dad jointly decide to use actions  $m_3$  and  $d_1$ , respectively, then neither of them sees to that the baby cries individually, but they see to it as a group. However, if they made these choices independently, without knowing each other’s action, then, even as a group, they perhaps should not be held responsible for the outcome. Similarly, in the International Salsa Competition example (Example 3), Bob, Chuck, and Dan (as a group) see to the team being disqualified if they coordinated their actions. For instance, if they spent the morning in a conversation at the time when they had to depart for the competition. We plan to explore these questions in our future work.

## References

- [1] H. G. Frankfurt, Alternate possibilities and moral responsibility, *The Journal of Philosophy* 66 (23) (1969) 829–839. doi:10.2307/2023833.
- [2] D. Widerker, *Moral responsibility and alternative possibilities: Essays on the importance of alternative possibilities*, Routledge, 2017.
- [3] F. Cushman, Deconstructing intent to reconstruct morality, *Current Opinion in Psychology* 6 (2015) 97–103.
- [4] D. Lewis, *Counterfactuals*, John Wiley & Sons, 2013.
- [5] J. Y. Halpern, *Actual causality*, MIT Press, 2016.
- [6] V. Batusov, M. Soutchanski, Situation calculus semantics for actual causality, in: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.
- [7] P. Naumov, J. Tao, Blameworthiness in strategic games, in: *Proceedings of Thirty-third AAAI Conference on Artificial Intelligence (AAAI-19)*, 2019.

- [8] P. Naumov, J. Tao, Blameworthiness in security games, in: Proceedings of Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20), 2020.
- [9] N. Belnap, M. Perloff, Seeing to it that: A canonical form for agentives, in: Knowledge representation and defeasible reasoning, Springer, 1990, pp. 167–190.
- [10] J. F. Horty, Agency and deontic logic, Oxford University Press, 2001.
- [11] J. F. Horty, N. Belnap, The deliberative STIT: A study of action, omission, ability, and obligation, *Journal of Philosophical Logic* 24 (6) (1995) 583–644.
- [12] J. Horty, E. Pacuit, Action types in STIT semantics, *The Review of Symbolic Logic* 10 (4) (2017) 617–637.
- [13] G. K. Olkhovikov, H. Wansing, Inference as doxastic agency. part i: The basics of justification STIT logic, *Studia Logica* 107 (1) (2019) 167–194.
- [14] D. Lewis, Causation as influence, *The Journal of Philosophy* 97 (4) (2000) 182–197.
- [15] J. Y. Halpern, A modification of the Halpern-Pearl definition of causality, in: Proceedings of the 24th International Joint Conference on Artificial Intelligence, 2015, pp. 3022–3033.
- [16] R. Zultan, T. Gerstenberg, D. A. Lagnado, Finding fault: causality and counterfactuals in group attributions., *Cognition* 125 (3) (2012) 429–440.
- [17] Aristotle, *The Nicomachean Ethics*, 10th Edition, Kegan Paul, Trench, Trübner & Co., 1906, translated by F.H. Peters.
- [18] A. L. Institute, 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, 1985 Print.
- [19] V. Yazdanpanah, M. Dastani, W. Jamroga, N. Alechina, B. Logan, Strategic responsibility under imperfect information, in: Proceedings

- of the 18th International Conference on Autonomous Agents and MultiAgent Systems, International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 592–600.
- [20] P. Naumov, J. Tao, An epistemic logic of blameworthiness, *Artificial Intelligence* 283, 103269 (June 2020).
- [21] P. Naumov, J. Tao, Duty to warn in strategic games, in: A. E. F. Seghrouchni, G. Sukthankar, B. An, N. Yorke-Smith (Eds.), *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20, Auckland, New Zealand, May 9-13, 2020, International Foundation for Autonomous Agents and Multiagent Systems, 2020*, pp. 904–912.
- [22] E. Lorini, D. Longin, E. Mayor, A logical analysis of responsibility attribution: emotions, individuals and collectives, *Journal of Logic and Computation* 24 (6) (2014) 1313–1339.
- [23] E. F. Witsell, D. D. Eisenhower, Procedure for military execution, u.S. Department of the Army Pamphlet No. 27-4, December 9th. [https://www.loc.gov/rr/frd/Military\\_Law/pdf/procedure\\_dec-1947.pdf](https://www.loc.gov/rr/frd/Military_Law/pdf/procedure_dec-1947.pdf) (December 1947).
- [24] M. Xu, Axioms for deliberative STIT, *Journal of Philosophical Logic* 27 (5) (1998) 505–552.
- [25] P. Balbiani, A. Herzig, N. Troquard, Alternative axiomatics and complexity of deliberative stit theories, *Journal of Philosophical Logic* 37 (4) (2008) 387–406.
- [26] J. Broersen, Deontic epistemic STIT logic distinguishing modes of mens rea, *Journal of Applied Logic* 9 (2) (2011) 137–152.
- [27] P. Naumov, J. Tao, Two forms of responsibility in strategic games, in: *30th International Joint Conference on Artificial Intelligence (IJCAI-21)*, 2021.
- [28] R. Berthon, B. Maubert, A. Murano, Decidability results for  $atl^*$  with imperfect information and perfect recall, in: *Proceedings of the 16th*

- Conference on Autonomous Agents and MultiAgent Systems, International Foundation for Autonomous Agents and Multiagent Systems, 2017, pp. 1250–1258.
- [29] R. Berthon, B. Maubert, A. Murano, S. Rubin, M. Y. Vardi, Strategy logic with imperfect information, in: Logic in Computer Science (LICS), 2017 32nd Annual ACM/IEEE Symposium on, IEEE, 2017, pp. 1–12.
  - [30] P. Naumov, J. Tao, Coalition power in epistemic transition systems, in: Proceedings of the 2017 International Conference on Autonomous Agents and Multiagent Systems (AAMAS), 2017, pp. 723–731.
  - [31] P. Naumov, J. Tao, Together we know how to achieve: An epistemic logic of know-how, *Artificial Intelligence* 262 (2018) 279 – 300. doi:<https://doi.org/10.1016/j.artint.2018.06.007>.
  - [32] P. Naumov, Y. Yuan, Intelligence in strategic games, *Journal of Artificial Intelligence Research* 71 (2021) 521–556.
  - [33] T. Ågotnes, N. Alechina, Coalition logic with individual, distributed and common knowledge, *Journal of Logic and Computation* 29 (2019) 1041–1069. doi:[10.1093/logcom/exv085](https://doi.org/10.1093/logcom/exv085).
  - [34] T. Ågotnes, Action and knowledge in alternating-time temporal logic, *Synthese* 149 (2) (2006) 375–407.
  - [35] J. Broersen, A. Herzig, N. Troquard, A STIT-extension of ATL, in: *European Workshop on Logics in Artificial Intelligence*, Springer, 2006, pp. 69–81.
  - [36] A. Herzig, N. Troquard, Knowing how to play: uniform choices in logics of agency, in: *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, 2006, pp. 209–216.
  - [37] J. Broersen, A. Herzig, N. Troquard, A normal simulation of coalition logic and an epistemic extension, in: *Proceedings of the 11th conference on Theoretical aspects of rationality and knowledge*, ACM, 2007, pp. 92–101.
  - [38] M. Pauly, Logic for social software, Ph.D. thesis, Institute for Logic, Language, and Computation (2001).

- [39] M. Pauly, A modal logic for coalitional power in games, *Journal of Logic and Computation* 12 (1) (2002) 149–166. doi:10.1093/logcom/12.1.149.
- [40] R. Alur, T. A. Henzinger, O. Kupferman, Alternating-time temporal logic, *Journal of the ACM* 49 (5) (2002) 672–713. doi:10.1145/585265.585270.
- [41] D. Walther, W. van der Hoek, M. Wooldridge, Alternating-time temporal logic with explicit strategies, in: *Proceedings of the 11th conference on Theoretical aspects of rationality and knowledge*, ACM, 2007, pp. 269–278.
- [42] T. De Lima, L. Royakkers, F. Dignum, A logic for reasoning about responsibility, *Logic Journal of IGPL* 18 (1) (2010) 99–117.
- [43] K. Deuser, P. Naumov, Strategic knowledge acquisition, *ACM Transactions on Computational Logic (TOCL)* 22 (3) (2021) 1–18.
- [44] W. Jamroga, W. van der Hoek, Agents that know how to play, *Fundamenta Informaticae* 63 (2-3) (2004) 185–219.
- [45] R. Fervari, A. Herzig, Y. Li, Y. Wang, Strategically knowing how, in: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 1031–1038. doi:10.24963/ijcai.2017/143.
- [46] R. Fervari, F. R. Velázquez-Quesada, Y. Wang, Bisimulations for knowing how logics, *The Review of Symbolic Logic* (2021) 1–37.
- [47] G. H. von Wright, *Norm and Action: A Logical Enquiry*, Routledge & Kegan Paul, 1963.
- [48] N. Belnap, M. Perloff, Seeing to it that: a canonical form for agentives, *Theoria* 54 (3) (1988) 175–199.
- [49] E. Mendelson, *Introduction to mathematical logic*, CRC press, 2009.

## Appendix A. Auxiliary Lemmas

**Lemma 8** *If  $X, \varphi \vdash \psi$ , then  $X \vdash \varphi \rightarrow \psi$ .*

PROOF. Suppose that sequence  $\psi_1, \dots, \psi_n$  is a proof from set  $X \cup \{\varphi\}$  and the theorems of our logical system that uses the Modus Ponens inference rule only. In other words, for each  $k \leq n$ , either

1.  $\vdash \psi_k$ , or
2.  $\psi_k \in X$ , or
3.  $\psi_k$  is equal to  $\varphi$ , or
4. there are  $i, j < k$  such that formula  $\psi_j$  is equal to  $\psi_i \rightarrow \psi_k$ .

It suffices to show that  $X \vdash \varphi \rightarrow \psi_k$  for each  $k \leq n$ . We prove this by induction on  $k$  through considering the four cases above separately.

**Case I:**  $\vdash \psi_k$ . Note that  $\psi_k \rightarrow (\varphi \rightarrow \psi_k)$  is a propositional tautology, and thus, is an axiom of our logical system. Hence,  $\vdash \varphi \rightarrow \psi_k$  by the Modus Ponens inference rule. Therefore,  $X \vdash \varphi \rightarrow \psi_k$ .

**Case II:**  $\psi_k \in X$ . Then, similar to the previous case,  $X \vdash \varphi \rightarrow \psi_k$ .

**Case III:** formula  $\psi_k$  is equal to  $\varphi$ . Thus,  $\varphi \rightarrow \psi_k$  is a propositional tautology. Therefore,  $X \vdash \varphi \rightarrow \psi_k$ .

**Case IV:** formula  $\psi_j$  is equal to  $\psi_i \rightarrow \psi_k$  for some  $i, j < k$ . Thus, by the induction hypothesis,  $X \vdash \varphi \rightarrow \psi_i$  and  $X \vdash \varphi \rightarrow (\psi_i \rightarrow \psi_k)$ . Note that formula  $(\varphi \rightarrow \psi_i) \rightarrow ((\varphi \rightarrow (\psi_i \rightarrow \psi_k)) \rightarrow (\varphi \rightarrow \psi_k))$  is a propositional tautology. Therefore,  $X \vdash \varphi \rightarrow \psi_k$  by applying the Modus Ponens inference rule twice.  $\square$

**Lemma 9**  $\vdash \mathbf{K}_a\varphi \rightarrow \mathbf{K}_a\mathbf{K}_a\varphi$ .

PROOF. Note that formula  $\mathbf{K}_a\neg\mathbf{K}_a\varphi \rightarrow \neg\mathbf{K}_a\varphi$  is an instance of the Truth axiom. Thus,  $\vdash \mathbf{K}_a\varphi \rightarrow \neg\mathbf{K}_a\neg\mathbf{K}_a\varphi$  by the law of contrapositive in propositional logic. Hence, taking into account the following instance of the Negative Introspection axiom  $\neg\mathbf{K}_a\neg\mathbf{K}_a\varphi \rightarrow \mathbf{K}_a\neg\mathbf{K}_a\neg\mathbf{K}_a\varphi$ , one can conclude that

$$\vdash \mathbf{K}_a\varphi \rightarrow \mathbf{K}_a\neg\mathbf{K}_a\neg\mathbf{K}_a\varphi. \quad (\text{A.1})$$

At the same time,  $\neg\mathbf{K}_a\varphi \rightarrow \mathbf{K}_a\neg\mathbf{K}_a\varphi$  is an instance of the Negative Introspection axiom. Thus,  $\vdash \neg\mathbf{K}_a\neg\mathbf{K}_a\varphi \rightarrow \mathbf{K}_a\varphi$  by contraposition. Hence,  $\vdash \mathbf{K}_a(\neg\mathbf{K}_a\neg\mathbf{K}_a\varphi \rightarrow \mathbf{K}_a\varphi)$  by the Necessitation inference rule. Thus,

$$\vdash \mathbf{K}_a\neg\mathbf{K}_a\neg\mathbf{K}_a\varphi \rightarrow \mathbf{K}_a\mathbf{K}_a\varphi$$

by the Distributivity axiom and the Modus Ponens inference rule. The last statement, together with statement (A.1), implies the statement of the lemma by the laws of propositional reasoning.  $\square$

**Lemma 10** *If  $\varphi_1, \dots, \varphi_n \vdash \psi$ , then  $\mathbf{K}_a\varphi_1, \dots, \mathbf{K}_a\varphi_n \vdash \mathbf{K}_a\psi$ .*

PROOF. By Lemma 8 applied  $n$  times, the assumption  $\varphi_1, \dots, \varphi_n \vdash \psi$  implies that  $\vdash \varphi_1 \rightarrow (\varphi_2 \rightarrow \dots (\varphi_n \rightarrow \psi) \dots)$ . Thus,

$$\vdash \mathbf{K}_a(\varphi_1 \rightarrow (\varphi_2 \rightarrow \dots (\varphi_n \rightarrow \psi) \dots))$$

by the Necessitation inference rule. Hence, by the Distributivity axiom and the Modus Ponens inference rule,

$$\vdash \mathbf{K}_a\varphi_1 \rightarrow \mathbf{K}_a(\varphi_2 \rightarrow \dots (\varphi_n \rightarrow \psi) \dots).$$

Then,  $\mathbf{K}_a\varphi_1 \vdash \mathbf{K}_a(\varphi_2 \rightarrow \dots (\varphi_n \rightarrow \psi) \dots)$ , again by the Modus Ponens inference rule. Therefore,  $\mathbf{K}_a\varphi_1, \dots, \mathbf{K}_a\varphi_n \vdash \mathbf{K}_a\psi$  by applying the previous steps  $(n - 1)$  more times.  $\square$

**Lemma 11**  $\vdash \mathbf{K}_a\varphi_1 \wedge \dots \wedge \mathbf{K}_a\varphi_n \leftrightarrow \mathbf{K}_a(\varphi_1 \wedge \dots \wedge \varphi_n)$ , if  $n \geq 0$ .

PROOF. Note that  $\varphi_1 \rightarrow (\varphi_2 \rightarrow \dots (\varphi_n \rightarrow \varphi_1 \wedge \dots \wedge \varphi_n) \dots)$  is a propositional tautology. Thus,  $\varphi_1, \dots, \varphi_n \vdash \varphi_1 \wedge \dots \wedge \varphi_n$  by applying the Modus Ponens inference rule  $n$  times. Hence,  $\mathbf{K}_a\varphi_1, \dots, \mathbf{K}_a\varphi_n \vdash \mathbf{K}_a(\varphi_1 \wedge \dots \wedge \varphi_n)$  by Lemma 10. Therefore,  $\vdash \mathbf{K}_a\varphi_1 \wedge \dots \wedge \mathbf{K}_a\varphi_n \rightarrow \mathbf{K}_a(\varphi_1 \wedge \dots \wedge \varphi_n)$  by propositional reasoning using Lemma 8.

To prove the implication in the other direction, consider any  $i \leq n$ . Then formula  $\varphi_1 \wedge \dots \wedge \varphi_n \rightarrow \varphi_i$  is a propositional tautology. Thus, by the Necessitation inference rule,  $\vdash \mathbf{K}_a(\varphi_1 \wedge \dots \wedge \varphi_n \rightarrow \varphi_i)$ . Hence, by the Distributivity axiom and the Modus Ponens inference rule,  $\vdash \mathbf{K}_a(\varphi_1 \wedge \dots \wedge \varphi_n) \rightarrow \mathbf{K}_a\varphi_i$  for each  $i \leq n$ . Therefore,  $\vdash \mathbf{K}_a(\varphi_1 \wedge \dots \wedge \varphi_n) \rightarrow \mathbf{K}_a\varphi_1 \wedge \dots \wedge \mathbf{K}_a\varphi_n$  by propositional reasoning.  $\square$