Two Forms of Responsibility in Strategic Games

Pavel Naumov[1] and **Jia Tao**[2]
[1]King's College
[2]Lafayette College
pgn2@cornell.edu, taoj@lafayette.edu

Abstract

The paper studies two forms of responsibility, seeing to it and being blamable, in the setting of strategic games with imperfect information. The paper shows that being blamable is definable through seeing to it, but not the other way around. In addition, it proposes a bimodal logical system that describes the interplay between the seeing to it modality and the individual knowledge modality.

1 Introduction

In this paper 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" [Frankfurt, 1969]. The principle of alternate possibilities is widely discussed in the literature [Widerker, 2017]. 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" [Cushman, 2015], is also used to define causality [Lewis, 2013; Halpern, 2016; Batusov and Soutchanski, 2018]. For the sake of clarity, in this paper we refer to all versions of responsibility based on the principle of alternate possibilities as *blameworthiness*. Formal logical systems for reasoning about blameworthiness in strategic and security games are proposed in [Naumov and Tao, 2019a] and [Naumov and Tao, 2020a] respectively.

The other approach is to hold a person responsible for the outcome if the person *sees to it that it happens*. In this paper 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 [Belnap and Perloff, 1990; Horty, 2001; Horty and Belnap, 1995; Horty and Pacuit, 2017; Olkhovikov and Wansing, 2018].

The rest of this paper is organized as follows. First, we illustrate the difference between the two forms of responsibility on an example. Then, we discuss how these two forms of responsibility could be defined in imperfect information setting. In Section 4 we formally define games with imperfect information. Section 5 defines modalities that capture seeing to it and blameworthiness. In the section that follows, we compare our semantics with epistemic XSTIT frames. Section 7 and Section 8 contain the two main results of this paper. We show that the blameable modality is definable through the seeing to it modality and that seeing to it is not definable through blameable even using the *ex ante* (before the action) knowledge modality. Section 9 discusses the future work and concludes.

2 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 games as the left and the right game. In these games, 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 paper 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 the baby cries.

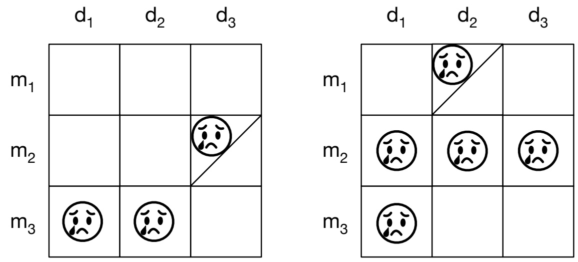

Figure 1: If baby cries under action profile (m_2, d_3) in the left game, the mom is *blamable* for baby crying. Under the same action profile in the right game, she *sees to it* that baby cries.

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 blamable for the baby crying because 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 mom is not blamable for it because in the right game she has no unilateral action that would prevent the baby from crying.

However, 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 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 baby crying.

Note that statement $2 + 2 = 4$ is true no matter what action the mom chooses. Thus, one can say that the mom sees to it that $2 + 2 = 4$. This is the approach taken in 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 leads to it. Horty and Belnap refer to seeing to it in the presence of such an alternative action as seeing to it "deliberately" [Horty and Belnap, 1995]. Since the focus of our paper is on two forms of 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.

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:

> *. . . 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.* *[Aristotle, 1906]*

In a legal setting, the 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 [Institute, 1985 Print].

If one considers games with imperfect information instead of games with perfect information that we discussed in the last section, then the definitions of both forms of responsibility must be adjusted to incorporate knowledge. In the case of blameworthiness, it is natural to require that the agent could be blamed 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 [Yazdanpanah *et al.*, 2019; Naumov and Tao, 2020c; Naumov and Tao, 2020b]. In the case of seeing to it, it is natural to require that not only should the agent's action unavoidably leads to the outcome, but the agent also must *interim* (at the time of the action) know this.

To illustrate the two forms of responsibility in 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 blameable for the death and sees to it. Indeed, the shooter is *blameable* for the death of the prisoner 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 also sees to the death of the prisoner 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 would not have to 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 could be blamed* for the death *individually* [1]. At the same time, each of the agents knows interim that pulling the trigger while aiming at the prisoner will unavoidably result in the 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 *sees to the death* of the prisoner.

In some cases, one or more members of the firing squad are issued a weapon containing a blank cartridge[2], 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 *blamable* 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 have a strategy to prevent the death unilaterally. In the "conscience bullet" setting, *none of the agents sees 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 blamable for the death and also the shooter sees to the death. If an execution by a firing squad is without a "conscience bullet", then none of the squad members is blameable, but each of them sees to the death. If at least one "conscience bullet" is used, none of the members is blamable for the death and none sees to the death.

4 Imperfect Information Strategic Games

In this section we give a formal definition of games with imperfect information used throughout the rest of this paper. Throughout the paper we assume a fixed set of propositional variables and a fixed set of agents $\mathcal{A}$.

[1] All shooters could be blamed together as a group under the coalitional blameworthiness definition in [Naumov and Tao, 2019b].

[2] "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." [Witsell and Eisenhower, 1947, p.5]

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

1. *I is a 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. *Δ_a^α is a nonempty set of "actions" for each agent $a \in \mathcal{A}$ and each state $\alpha \in I$, where for each agent $a \in \mathcal{A}$ and all states $\alpha, \alpha' \in I$, if $\alpha \sim_a \alpha'$, then $\Delta_a^\alpha = \Delta_a^{\alpha'}$,*
4. *Ω is a set of "outcomes",*
5. *P is a set of triples (α, δ, ω), called "plays", where $\alpha \in I$ is an initial state, $\omega \in \Omega$ is an outcome, and*
    - *(a) function δ, called an "action profile in state α", 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 δ in state α, 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.*

The indistinguishability equivalence relation between initial states $\sim_a$ captures the *ex ante* knowledge of an agent a or the knowledge of the agent before the transition takes place. Knowledge defined through equivalence relation on outcomes is usually called the *ex post* knowledge. We do not include the *ex post* knowledge into our system because it is not relevant to responsibility.

The assumption of item 3 in Definition 1 that $\Delta_a^\alpha = \Delta_a^{\alpha'}$ when $\alpha \sim_a \alpha'$ states that an agent has the same set of actions in all indistinguishable states. In other words, the set of available actions is *known* to the agent.

In Definition 1, we distinguish the set of initial states I from the set of outcomes Ω. This is done for convenience only. In particular, we allow I and Ω to be the same set.

The set of plays P captures the mechanism or the set of "rules" of the game. It specifies into which outcome the game can transition from a given initial state under a given action profile. We allow the mechanism to be nondeterministic.

Finally, unlike many other logical systems, we use propositional variables to represent not statements about outcomes, but, more generally, statements about plays. We further discuss the reason for this choice in the next section. Formally, this is captured through the value of $\pi(p)$ being a set of plays rather than a set of outcomes.

5 Syntax and Semantics

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

$$\varphi := p \mid \neg\varphi \mid \varphi \to \varphi \mid \mathsf{K}_a\varphi \mid \mathsf{ST}_a\varphi \mid \mathsf{B}_a\varphi,$$

where p is a propositional variable and $a \in \mathcal{A}$ is an agent. We read $\mathsf{K}_a\varphi$ as "agent a knows ex ante that statement φ will be true", $\mathsf{B}_a\varphi$ as "agent a is blameable for φ", and $\mathsf{ST}_a\varphi$ as "agent a sees to φ". By Φ^{B} we denote the fragment of the language $\Phi^{\mathsf{ST},\mathsf{B}}$ that does not include modality ST. Similarly, by Φ^{ST} we denote the fragment of $\Phi^{\mathsf{ST},\mathsf{B}}$ without modality B.

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

1. *$(\alpha, \delta, \omega) \Vdash p$, if $(\alpha, \delta, \omega) \in \pi(p)$,*
2. *$(\alpha, \delta, \omega) \Vdash \neg\varphi$, if $(\alpha, \delta, \omega) \nVdash \varphi$,*
3. *$(\alpha, \delta, \omega) \Vdash \varphi \to \psi$, if $(\alpha, \delta, \omega) \nVdash \varphi$ or $(\alpha, \delta, \omega) \Vdash \psi$,*
4. *$(\alpha, \delta, \omega) \Vdash \mathsf{K}_a\varphi$, if $(\alpha', \delta', \omega') \Vdash \varphi$ for each play $(\alpha', \delta', \omega') \in P$ such that $\alpha \sim_a \alpha'$,*
5. *$(\alpha, \delta, \omega) \Vdash \mathsf{ST}_a\varphi$, 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) $(\alpha', \delta', \omega') \nVdash \varphi$ for some play $(\alpha', \delta', \omega') \in P$ such that $\alpha \sim_a \alpha'$.*
6. *$(\alpha, \delta, \omega) \Vdash \mathsf{B}_a\varphi$ 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 $(\alpha', \delta', \omega') \nVdash \varphi$.*

The formal semantics in Definition 2 specifies satisfaction $\Vdash$ as a relation between a play (α, δ, ω) and a formula φ. 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" is a property not of a state, but rather of a transition between states[3]. Indeed, consider the game depicted in Figure 2 (left). This game has a single initial state α and two outcome states ω_1 and ω_2. The only agent *Alice* of this game has two possible actions in state α: action a and action b. Action a deterministically transitions the game into outcome ω_1 and action b nondeterministically leads to either outcome ω_1 or outcome ω_2. Let statement p stand for "game ends with outcome ω_1". Note that if Alice employs action a, then she knows at the time of the action, that statement p will be true and she also knows that p is not unavoidable if she chooses action b instead. Thus, by item 5 of Definition 2, she sees to it that p is true:

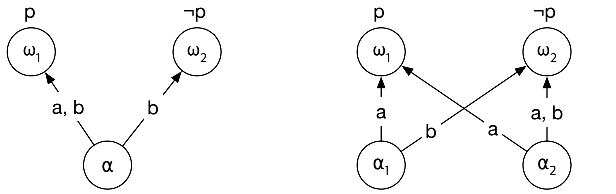

Figure 2: Two Games.

$$(\alpha, a, \omega_1) \Vdash \mathsf{ST}_{Alice}p.$$

At the same time, if the game ends with the same outcome ω_1, but Alice uses action b instead of a, then she does see to p because condition 5(a) of Definition 2 is not satisfied:

$$(\alpha, b, \omega_1) \nVdash \mathsf{ST}_{Alice}p.$$

Hence, whether an agent sees to something depends not only on the outcome, but also on the actions of the agent that lead to that outcome.

[3] The same is true about the blameworthiness modality as well.

The game depicted in Figure 2 (right) illustrates that seeing to something also depends on the initial state:

$$(\alpha_1, a, \omega_1) \Vdash \mathsf{ST}_{Alice}p, \qquad (\alpha_2, a, \omega_1) \nVdash \mathsf{ST}_{Alice}p.$$

Thus, a statement of the form $\mathsf{ST}_a\varphi$ depends not only on the action profile, but also on the initial state. Because formulae in our languages use this modality, we define satisfaction relation $\Vdash$ as a relation between plays[4] and formulae.

Next, let us turn to item 5(b) of the above definition. 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 Delibarative STIT [Horty and Belnap, 1995]. 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 φ is unavoidable,
2. φ is avoidable,
3. agent a knows that φ is avoidable.

Out of these alternatives, 1 is the weakest and 3 is the strongest. We believe that condition 1 is the best to capture the "guilty mind" aspect of responsibility. 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. Item 5(b) of Definition 2 formally captures alternative 1.

6 Games vs Epistemic XSTIT Frames

The original STIT logic does not include a knowledge modality and, thus, its standard semantics does not contain an equivalence relation on states. Broersen, Herzig, and Troquard propose a STIT-extension of ATL and also define an epistemic version of this extension [Broersen *et al.*, 2006]. In [Broersen *et al.*, 2007] they discuss a possibility of adding knowledge modality to NCL, a version of STIT, and claim completeness without a proof. The most detailed account of their epistemic extension of STIT and its semantics in terms of epistemic XSTIT frames is given in [Broersen, 2011]. In this section we compare this work with ours.

The standard semantics of STIT is defined with respect to dynamic states or pairs of a history and a state in the history. History roughly corresponds to our notion of a play. As we mentioned earlier, there are three forms of knowledge that can be considered for any action: ex ante (before action), interim (at the moment of action), and ex post (after outcome is known). Ex ante knowledge is defined in terms of indistinguishable initial states, interim knowledge – in terms of indistinguishable initial states and equal actions, and ex post knowledge – in terms of indistinguishable initial state, equal actions, and indistinguishable outcomes. Item 5(a) of Definition 2 refers to interim knowledge because it requires states α and α' be indistinguishable and actions $\delta(a)$ and $\delta'(a)$ be the same. Item 5(b) of Definition 2 refers to ex ante knowledge because it only requires states to be indistinguishable, but allows actions $\delta(a)$ and $\delta'(a)$ to be different.

Epistemic XSTIT frames define knowledge through an indistinguishably relation on dynamic states. [Broersen, 2011] states "if we let uncertainty range over dynamic states, as for the present logic, we can talk about knowledge of what agents are doing". However, epistemic XSTIT frames do not include any condition that connects relation R_a, representing agent's actions, with relation $\sim_a$, representing agent's knowledge. As a result, the knowledge modality could be potentially interpreted as capturing any of the three forms of knowledge we discussed above. In fact, because histories include future states, the same knowledge modality could also be interpreted as capturing knowledge of a fortune-teller who can see the future. [Broersen, 2011, p.147] uses this very abstract knowledge modality to express what is supposed to be an equivalent version of our modality ST:

$$[a\ \mathsf{dxstit}]''\varphi \equiv_{def} \mathsf{K}_a[a\ \mathsf{xstit}]\varphi \wedge \neg\mathsf{K}_a \Box \mathsf{X}\varphi. \tag{1}$$

The first conjunct on the right-hand-side of the above formula states that the agent knows (modality K) that her action would unavoidably lead (modality $[a\ \mathsf{xstit}]$) to statement φ being true. This conjunct intends to express condition 5(a) of Definition 2. Since the agent's action is not fixed at ex ante time, modality K_a must refer to interim knowledge, just like our condition 5(a). At the same time, the second conjunct states that the agent does not know (modality K) that her action unavoidably lead (the combination of modalities $\Box\mathsf{X}$) to statement φ being true. This statement intends to express condition 5(b) of Definition 2. Note that knowledge here should be ex ante because once the agent's action is fixed, statement φ is guaranteed to be true (as per first conjunct). Therefore, in order to faithfully capture the intended meaning, the two instances of modality K on the right-hand-side of equation (1) must refer to two different forms of knowledge: ex ante and interim, which is not the case for epistemic XSTIT frames.

A version of epistemic STIT logic that distinguishes ex ante, interim, and ex post forms of knowledge is proposed in [Lorini *et al.*, 2014]. Their notion of *active agentive responsibility* is the one captured by ST modality in the current paper. They study the decidability of the satisfiability problem, but do not propose an axiomatization. In this paper we distinguish the two forms of knowledge in items 5(a) and 5(b) of Definition 2 and we use modality K to represent ex ante knowledge. Our main technical results are the definability of modality B through ST and a sound and complete logical system capturing the properties of modality ST.

7 Definability of B through ST

In this section we show that blameworthiness modality B is expressible through seeing to it modality ST. In the next section, we show that the opposite is false.

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

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

$$(\alpha, \delta, \omega) \Vdash \varphi \tag{2}$$

[4] We also include an outcome ω because we want to be able to reason about a more general class of *nondeterministic* games in which an outcome is not uniquely determined by the actions.

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

$$\alpha \sim_a \alpha' \wedge d = \delta'(a) \Rightarrow (\alpha', \delta', \omega') \nVdash \varphi. \tag{3}$$

Due to statement (2), it suffices to show that

$$(\alpha, \delta, \omega) \Vdash \mathsf{ST}_a \neg \mathsf{ST}_a \neg \varphi. \tag{4}$$

We prove this statement by verifying the two claims below.

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

PROOF OF CLAIM. Statement (2) implies that $(\alpha'', \delta'', \omega'') \nVdash \mathsf{ST}_a \neg\varphi$ by item 5 of Definition 2 because $\alpha \sim_a \alpha''$ and $\delta(a) = \delta''(a)$. Therefore, $(\alpha'', \delta'', \omega'') \Vdash \neg\mathsf{ST}_a \neg\varphi$ by item 2 of Definition 2. $\boxtimes$

Let δ_0 be any action profile in state α such that $\delta_0(a) = d$. Such an action profile exists because the domain of actions of each agent in state α is not empty by item 3 of Definition 1. Then, by condition 5(b) of Definition 1, there must exist at least one outcome ω_0 such that $(\alpha, \delta_0, \omega_0) \in P$.

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

PROOF OF CLAIM. Recall that $\delta_0(a) = d$ by the choice of action profile δ_0. Thus, $(\alpha', \delta', \omega') \nVdash \varphi$ for each play $(\alpha', \delta', \omega') \in P$ such that $\alpha \sim_a \alpha'$ and $\delta_0(a) = \delta'(a)$ by statement (3). At the same time $(\alpha, \delta, \omega) \Vdash \varphi$ by statement (2). Therefore, $(\alpha, \delta_0, \omega_0) \Vdash \mathsf{ST}_a \neg\varphi$ by item 5 of Definition 2. $\boxtimes$

Claims 1 and 2 imply statement (4) by item 5 of Definition 2.

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

$$(\alpha, \delta, \omega) \Vdash \varphi \tag{5}$$

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

$$\alpha \sim_a \alpha' \tag{6}$$

and $(\alpha', \delta', \omega') \nVdash \neg\mathsf{ST}_a \neg\varphi$. Thus, $(\alpha', \delta', \omega') \Vdash \mathsf{ST}_a \neg\varphi$ by item 2 of Definition 2. Then, for each play $(\alpha'', \delta'', \omega'') \in P$,

$$\alpha' \sim_a \alpha'' \wedge \delta'(a) = \delta''(a) \Rightarrow (\alpha'', \delta'', \omega'') \Vdash \neg\varphi \tag{7}$$

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 (6). Then, by statement (6) and statement (7), for any play $(\alpha'', \delta'', \omega'') \in P$,

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

Therefore, $(\alpha, \delta, \omega) \Vdash \mathsf{B}_a \varphi$ by item 6 of Definition 2 and statement (5). $\boxtimes$

8 Undefinability of ST through B and K

As we have seen in Theorem 1, blameworthiness modality B could be defined through seeing to it modality ST. In this section we show that modality ST cannot be defined in language Φ^{B}. To prove this, we construct two games and define a common play for both games such that a formula $\varphi \in \Phi^{\mathsf{B}}$ 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 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.

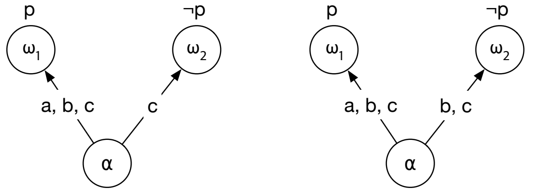

Figure 3: Two Games.

The two games are depicted in Figure 3. Each game has a single agent, Alice. In other words, $\mathcal{A} = \{Alice\}$. Both games have a single initial state α and two outcomes: ω_1 and ω_2. In both games, Alice has three actions (a, b, and c) in state α. In both games, propositional variable[5] p holds only on plays that result in outcome ω_1. In both games, action a leads to outcome ω_1 and action c leads (nondeterministically) to outcome ω_1 or outcome ω_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 3.

We refer to the two games from Figure 3 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 π_l and π_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 (α, b, ω_1) is the common play of these two games that we use to show the undefinability of modality ST in language Φ^{B}.

As mentioned above, Lemma 4 is a key step in the proof of the 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 $\omega \in \{\omega_1, \omega_2\}$.* $\boxtimes$

[5] We assume here that the language contains only one propositional variable. If the language contains more variables, satisfaction relation of all of them should be defined the same way as p.

Second, observe that plays (α, a, ω_1) and (α, b, ω_1) in the left game are indistinguishabile in language Φ^{B}.

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

The third lemma establishes that actions b and c in the right game are indistinguishabile in language Φ^{B}. The proof of this lemma is similar to the proof of Lemma 2.

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 $\varphi \in \Phi^{\mathsf{B}}$.* $\boxtimes$

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 on 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 $\varphi \in \Phi^{\mathsf{B}}$ and each play $(\alpha, \delta, \omega) \in P_l$.*

Proof. We prove the lemma by structural induction on formula φ. If φ 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 φ is a negation or an implication follows from the induction hypothesis and items 2 and 3 of Definition 2 respectively.

Suppose that formula φ has the form $\mathsf{K}_{Alice}\psi$. Recall that there is only one state that Alice cannot distinguish from state α – the state α itself.
$(\Rightarrow)$: Assume that $(\alpha, \delta, \omega) \Vdash_l \mathsf{K}_{Alice}\psi$. 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 $(\alpha, \delta', \omega') \in P_r$ because $P_r = P_l \cup \{(\alpha, b, \omega_2)\}$ and due to Lemma 3.
$(\Leftarrow)$: Suppose that $(\alpha, \delta, \omega) \Vdash_r \mathsf{K}_{Alice}\psi$. 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, $(\alpha, \delta, \omega) \Vdash_l \mathsf{K}_{Alice}\psi$ by item 4 of Definition 2.

Finally, suppose that formula φ has the form $\mathsf{B}_{Alice}\psi$.
$(\Rightarrow)$: Let $(\alpha, \delta, \omega) \Vdash_l \mathsf{B}_{Alice}\psi$ for some play $(\alpha, \delta, \omega) \in P_l$. Hence, by item 6 of Definition 2,

$$(\alpha, \delta, \omega) \Vdash_l \psi \tag{8}$$

and $\exists x \in \{a, b, c\}\ \forall(\alpha, x, \omega') \in P_l\ ((\alpha, x, \omega') \nVdash_l \psi)$. Thus, by Lemma 2,

$$\exists x \in \{a, c\}\ \forall(\alpha, x, \omega') \in P_l\ ((\alpha, x, \omega') \nVdash_l \psi).$$

Then, by the induction hypothesis,

$$\exists x \in \{a, c\}\ \forall(\alpha, x, \omega') \in P_l\ ((\alpha, x, \omega') \nVdash_r \psi).$$

Hence, by Lemma 1,

$$\exists x \in \{a, c\}\ \forall(\alpha, x, \omega') \in P_r\ ((\alpha, x, \omega') \nVdash_r \psi).$$

In addition, $(\alpha, \delta, \omega) \Vdash_r \psi$ also by the induction hypothesis using statement (8). Therefore, $(\alpha, \delta, \omega) \Vdash_r \mathsf{B}_{Alice}\psi$ by item 6 of Definition 2.
$(\Leftarrow)$: Suppose $(\alpha, \delta, \omega) \Vdash_r \mathsf{B}_{Alice}\psi$. Thus, by item 6 of Definition 2,

$$(\alpha, \delta, \omega) \Vdash_r \psi \tag{9}$$

and $\exists x \in \{a, b, c\}\ \forall(\alpha, x, \omega') \in P_r\ ((\alpha, x, \omega') \nVdash_r \psi)$. Then, by Lemma 3,

$$\exists x \in \{a, c\}\ \forall(\alpha, x, \omega') \in P_r\ ((\alpha, x, \omega') \nVdash_r \psi).$$

Hence, by Lemma 1,

$$\exists x \in \{a, c\}\ \forall(\alpha, x, \omega') \in P_l\ ((\alpha, x, \omega') \nVdash_r \psi).$$

Thus, by the induction hypothesis,

$$\exists x \in \{a, c\}\ \forall(\alpha, x, \omega') \in P_l\ ((\alpha, x, \omega') \nVdash_l \psi).$$

In addition, $(\alpha, \delta, \omega) \Vdash_l \psi$ also by the induction hypothesis using statement (9). Therefore, $(\alpha, \delta, \omega) \Vdash_l \mathsf{B}_{Alice}\psi$ by item 6 of Definition 2. $\boxtimes$

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. *$(\alpha, b, \omega_1) \Vdash_l \mathsf{ST}_{Alice}p$.* $\boxtimes$

Lemma 6. *$(\alpha, b, \omega_1) \nVdash_r \mathsf{ST}_{Alice}p$.* $\boxtimes$

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

Theorem 2. *Modality ST is not definable in the language Φ^{B}.*

9 Conclusion

In this paper we study the ex ante knowledge modality and two responsibility modalities: "seeing to it" and "being blamable". We observed that "being blamable" could be defined through "seeing to it" and have shown that the converse is not true.

In the future, we plan to develop a complete logical system describing the interplay between modalities K and ST in language Φ^{ST}. Such a logical system would be different from the epistemic version of STIT described in [Broersen, 2011]. This is because our semantics is based on the strategic games with imperfect information semantics rather than properties of epistemic XSTIT frames. Furthermore, as we discussed in Section 6, we interpret K as ex ante knowledge instead of a very general notion of knowledge modeled by epistemic XSTIT frames. As a result, our system would contain new logical principles such as $\mathsf{K}_a\varphi \to \neg\mathsf{ST}_a\varphi$. This axiom is sound if modality K_a represents the ex ante knowledge, as in the proposed system, but is *not* sound, for example, if K_a represents the interim knowledge.

References

[Aristotle, 1906] Aristotle. *The Nicomachean Ethics*. Kegan Paul, Trench, Trübner & Co., 10th edition, 1906. translated by F.H. Peters.

[Batusov and Soutchanski, 2018] Vitaliy Batusov and Mikhail Soutchanski. Situation calculus semantics for actual causality. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

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

[Broersen *et al.*, 2006] Jan Broersen, Andreas Herzig, and Nicolas Troquard. A STIT-extension of ATL. In *European Workshop on Logics in Artificial Intelligence*, pages 69–81. Springer, 2006.

[Broersen *et al.*, 2007] Jan Broersen, Andreas Herzig, and Nicolas Troquard. A normal simulation of coalition logic and an epistemic extension. In *Proceedings of the 11th conference on Theoretical aspects of rationality and knowledge*, pages 92–101. ACM, 2007.

[Broersen, 2011] Jan Broersen. Deontic epistemic STIT logic distinguishing modes of mens rea. *Journal of Applied Logic*, 9(2):137–152, 2011.

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

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

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

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

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

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

[Institute, 1985 Print] American Law 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.

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

[Lorini *et al.*, 2014] Emiliano Lorini, Dominique Longin, and Eunate Mayor. A logical analysis of responsibility attribution: emotions, individuals and collectives. *Journal of Logic and Computation*, 24(6):1313–1339, 2014.

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

[Naumov and Tao, 2019b] Pavel Naumov and Jia Tao. Knowing-how under uncertainty. *Artificial Intelligence*, 276:41 – 56, 2019.

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

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

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

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

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

[Witsell and Eisenhower, 1947] Edward F. Witsell and Dwight D. Eisenhower. Procedure for military execution, December 1947. 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.

[Yazdanpanah *et al.*, 2019] Vahid Yazdanpanah, Mehdi Dastani, Wojciech Jamroga, Natasha Alechina, and Brian Logan. Strategic responsibility under imperfect information. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 592–600. International Foundation for Autonomous Agents and Multiagent Systems, 2019.