Blameworthiness in Security Games

Pavel Naumov
Department of Mathematical Sciences
Claremont McKenna College
Claremont, California 91711
pgn2@cornell.edu

Jia Tao
Department of Computer Science
Lafayette College
Easton, Pennsylvania 18042
taoj@lafayette.edu

Abstract

Security games are an example of a successful real-world application of game theory. The paper defines blameworthiness of the defender and the attacker in security games using the principle of alternative possibilities and provides a sound and complete logical system for reasoning about blameworthiness in such games. Two of the axioms of this system capture the asymmetry of information in security games.

Introduction

In this paper we study the properties of blameworthiness in security games (von Stackelberg 1934). Security games are used for canine airport patrol (Pita et al. 2008; Jain et al. 2010), airport passenger screening (Brown et al. 2016), protecting endangered animals and fish stocks (Fang, Stone, and Tambe 2015), U.S. Coast Guard port patrol (Sinha et al. 2018; An, Tambe, and Sinha 2016), and randomized deployment of U.S. air marshals (Sinha et al. 2018).

Defender \Attacker	Terminal 1	Terminal 2
Terminal 1	20	120
Terminal 2	200	16

Figure 1: Expected Human Losses in Security Game G_1.

As an example, consider a security game G_1 in which a defender is trying to protect two terminals in an airport from an attacker. Due to limited resources, the defender can patrol only one terminal at a given time. If the defender chooses to patrol Terminal 1 and the attacker chooses to attack Terminal 2, then the human losses at Terminal 2 are estimated at 120, see Figure 1. However, if the defender chooses to patrol Terminal 2 while the attacker still chooses to attack Terminal 2, then the expected number of the human losses at Terminal 2 is only 16, see Figure 1. Generally speaking, the goal of the defender is to minimize human losses, while the goal of the attacker is to maximize them. However, the utility functions in security games usually take into account not only the human losses, but also the cost to protect and to attack the target to the defender and the attacker respectively. Such a cost has to be converted to human lives using some factor, possibly different for the defender and the attacker. In game G_1, we assume that the cost of defending Terminal 1 and Terminal 2 is 8 and 4 respectively, while the cost of attacking these terminals is 12 and 8 respectively, see Figure 2. As a result, for example, if the defender chooses to patrol Terminal 1 and the attacker chooses to attack Terminal 2, then the payoff of the defender is $-120 - 8 = -128$ and the payoff of the attacker is $120 - 8 = 112$, see Figure 2.

Defender \Attacker	Terminal 1 (cost 12)	Terminal 2 (cost 8)
Terminal 1 (cost 8)	$-28, 8$	$-128, 112$
Terminal 2 (cost 4)	$-204, 188$	$-20, 8$

Figure 2: Utility Functions in Security Game G_1.

In real world examples of security games, the defender usually employs mixed strategies. For example, if the defender is using a strategy $75/25$, then he will spend 75% of the time in Terminal 1 and 25% of the time in Terminal 2. In practice, each morning the defender might get a randomly generated timetable that specifies at which terminal the defender should be at each time slot during the day (Jain et al. 2010). The distinctive feature of security games compared to strategic games is *the asymmetry of information* between the players: the attacker knows the strategy employed by the defender but not vice versa. For example, while planning the attack, the attacker might visit the airport multiple times and observe that the defender spends 75% of the time in Terminal 1 and 25% of the time in Terminal 2. Thus, the attacker will know the mixed strategy used by the defender, but she will not know the location of the defender at the moment she plans to arrive at the airport on the day of the attack.

For the sake of simplicity, we assume that in game G_1 the defender must choose between only three given mixed strategies: $75/25$, $50/50$, and $25/75$. Then, game G_1 can be described as an extensive form game depicted in Figure 3. The payoffs in this figure represent expected values of the utility functions. For example, suppose that the defender chooses the mixed strategy $75/25$ and the attacker chooses to attack Terminal 1. The pair $(75/25, T1)$ is called an *action profile* of game G_1. Under this action profile, the payoffs of the defender and the attacker are -28 and 8, respectively, with probability 75%, and they are -204 and 188, respectively, with probability 25%, see Figure 2. Thus, the

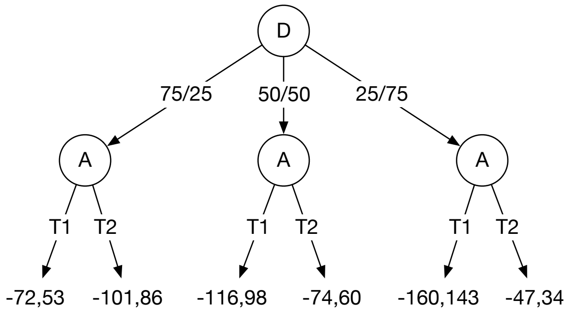

Figure 3: Security Game G_1 in Extensive Form.

expected payoff (or just "payoff") of the defender is

$$75\% \times (-28) + 25\% \times (-204) = -21 - 51 = -72$$

and of the attacker is

$$75\% \times 8 + 25\% \times 188 = 6 + 47 = 53.$$

Suppose that the defender chooses a strategy 50/50 and the attacker decides to target Terminal 2. Then, the attacker's payoff is 60, see Figure 3. We write this as

$$(50/50, T2) \Vdash \text{"The attacker's payoff is 60.".}$$

The attacker's mastermind might find this to be the attacker's fault and *blame* the attacker for the payoff not being at least 98. We capture the attacker's blameworthiness by

$$(50/50, T2) \Vdash \mathsf{A}(\text{"The attacker's payoff is less than 98."}),$$

where the blameworthiness modality $\mathsf{A}\varphi$ stands for "the attacker is blamable for φ". We define the blameworthiness using the well known Frankfurt's principle[1] of alternative possibilities: *an agent is blamable for φ if φ is true and the agent could have prevented φ* (Frankfurt 1969; Widerker 2017). In our case, the attacker, after learning that the defender's strategy is 50/50, could have targeted Terminal 1, which would increase her payoff to 98, see Figure 3. The principle of alternative possibilities, sometimes referred to as "counterfactual possibility" (Cushman 2015), is also used to define causality (Lewis 2013; Halpern 2016; Batusov and Soutchanski 2018).

Next, assume that the defender still chooses the strategy $50/50$, but the attacker decided to target Terminal 1. Under this action profile, the payoff of the attacker is 98, see Figure 3. Although the payoff is less than the attacker's payoff of 143 under the action profile $(25/75, T1)$, the attacker cannot be blamed for this:

$$(50/50, T1) \Vdash \neg\mathsf{A}(\text{"The attacker's payoff is less than 143."}),$$

because the attacker had no action in game G_1 to guarantee her payoff to be at least 143. At the same time, under the action profile $(25/75, T1)$, the defender is blameable for his payoff being less than -101:

$$(50/50, T1) \Vdash \mathsf{D}(\text{"The defender's payoff is less than } -101.\text{"}),$$

because the defender could have guaranteed his payoff to be at least -101 by choosing mixed strategy $75/25$, see Figure 3. Following the principle of alternative possibilities, the blameworthiness modality $\mathsf{D}\varphi$ stands for "statement φ is true and the defender had a strategy to prevent it".

In addition to the blameworthiness modalities A and D, in this paper we also consider an auxiliary necessity modality N. Statement $\mathsf{N}\varphi$ stands for "φ is true under each action profile of the given security game". For example,

$$(50/50, T1) \Vdash \mathsf{N}(\text{"The defender's payoff is negative."}),$$

because in game G_1 the defender's payoff is always negative. Surprisingly, as we show in Lemma 1, modality D can be expressed through modalities A and N:

$$\mathsf{D}\varphi \equiv \varphi \wedge \neg\mathsf{N}(\neg\varphi \to \mathsf{A}\neg\varphi).$$

At the same time, we believe that modality A cannot be expressed through modalities D and N, which reflects the *asymmetric* nature of security games.

In this paper we give a sound and complete axiomatization of the interplay between modalities A and N in security games. This work is related to our paper on blameworthiness in strategic games (2019). They proposed a sound and complete axiomatization of the interplay between the necessity modality N and the coalition blameworthiness modality B_C in strategic games. Their definition of the blameworthiness is also based on the principle of alternative possibilities. Namely, $\mathsf{B}_C\varphi$ stands for "statement φ is true and coalition (a set of agents) C had a strategy to prevent it". Thus, our modalities $\mathsf{A}\varphi$ and $\mathsf{D}\varphi$ correspond to their modalities $\mathsf{B}_{\{\text{attacker}\}}\varphi$ and $\mathsf{B}_{\{\text{defender}\}}\varphi$. In spite of this *syntactic similarity* between their and our works, the resulting axiomatic systems are quite different, which comes from the *semantic difference* between strategic games and security games. In security games, the attacker knows the defender's strategy while in a similar strategic game she would not. There are three aspects in which this work is different from (Naumov and Tao 2019):

1. As stated above, in security games modality D is expressible through modalities A and N, while in strategic games modality $\mathsf{B}_{\{\text{defender}\}}$ is not expressible through modalities $\mathsf{B}_{\{\text{attacker}\}}$ and N.
2. Two of our core axioms for modality A, the Conjunction axiom and the No Blame axiom capture the asymmetry of information in security games. They are not sound in strategic games. The Fairness axiom from (Naumov and Tao 2019) is not sound in our setting. We further discuss this in the Axioms section.
3. The proof of the completeness is using a completely different construction from the one used in (Naumov and Tao 2019). This is discussed in section Completeness.

Syntax and Semantics

In this paper we consider a fixed set of propositional variables Prop. The language Φ of our logical system is defined by the grammar: $\varphi := p \mid \neg\varphi \mid \varphi \to \varphi \mid \mathsf{N}\varphi \mid \mathsf{A}\varphi$.

As usual, we assume that connectives $\wedge$, $\vee$, and $\leftrightarrow$ are defined through connectives $\to$ and $\neg$ in the standard way. Next, we formally define security games (or just "games").

[1] This principle has many limitations that (Frankfurt 1969) discusses; for example, when a person is coerced into something.

Definition 1 *A game is a tuple $(\mathcal{D}, \{\mathcal{A}_d\}_{d\in\mathcal{D}}, \pi)$, where*

1. *set $\mathcal{D}$ is a set of actions of the defender,*
2. *non-empty set $\mathcal{A}_d$ is a set of actions of the attacker in response to the action $d \in \mathcal{D}$ of the defender,*
3. *valuation $\pi(p)$ of a propositional variable p is an arbitrary set of pairs (d, a) such that $d \in \mathcal{D}$ and $a \in \mathcal{A}_d$.*

In game G_1 from the introduction, the set of actions $\mathcal{D}$ of the defender is a three-element set $\{75/25, 50/50, 25/75\}$. For each action $d \in \mathcal{D}$ of the defender in this game, the set of responses $\mathcal{A}_d$ is the same two-element set $\{T1, T2\}$. Informally, $\pi(p)$ describes the set of action profiles (d, a) under which statement p is true.

The next definition is the core definition of our paper. Its item 5 defines blameworthiness of the attacker in security games using the principle of alternative possibilities (Frankfurt 1969; Widerker 2017): the attacker is blamable for statement φ under action profile (d, a) if φ is true under this profile and the attacker had an opportunity to prevent φ.

Definition 2 *For any action $d \in \mathcal{D}$ of the defender and any response action $a \in \mathcal{A}_d$ of the attacker in a game $(\mathcal{D}, \{\mathcal{A}_d\}_{d\in\mathcal{D}}, \pi)$ and any formula $\varphi \in \Phi$, the satisfiability relation $(d, a) \Vdash \varphi$ is defined recursively as follows:*

1. *$(d, a) \Vdash p$ if $(d, a) \in \pi(p)$, where $p \in \mathsf{Prop}$,*
2. *$(d, a) \Vdash \neg\varphi$ if $(d, a) \nVdash \varphi$,*
3. *$(d, a) \Vdash \varphi \to \psi$ if $(d, a) \nVdash \varphi$ or $(d, a) \Vdash \psi$,*
4. *$(d, a) \Vdash \mathsf{N}\varphi$ if $(d', a') \Vdash \varphi$ for each action $d' \in \mathcal{D}$ of the defender and each response action $a' \in \mathcal{A}_{d'}$ of the attacker,*
5. *$(d, a) \Vdash \mathsf{A}\varphi$ if $(d, a) \Vdash \varphi$ and there is a response action $a' \in \mathcal{A}_d$ of the attacker such that $(d, a') \nVdash \varphi$.*

As defined above, language Φ includes the attacker's blameworthiness modality A, but does not include the defender's blameworthiness modality D. If modality D is added to language Φ to form language Φ^+, then Definition 2 would need to be extended by an additional item:

6. *$(d, a) \Vdash \mathsf{D}\varphi$ if $(d, a) \Vdash \varphi$ and there is an action $d' \in \mathcal{D}$ of the defender such that for each response action $a' \in \mathcal{A}_{d'}$ of the attacker, $(d', a') \nVdash \varphi$.*

As mentioned in the introduction, we do not include modality D into language Φ because it is expressible through modalities A and N. Indeed, the following lemma holds for any formula $\varphi \in \Phi^+$:

Lemma 1 $(d, a) \Vdash \mathsf{D}\varphi$ *iff* $(d, a) \Vdash \varphi \wedge \neg\mathsf{N}(\neg\varphi \to \mathsf{A}\neg\varphi)$.

PROOF. $(\Rightarrow)$: Suppose that $(d, a) \nVdash \varphi \wedge \neg\mathsf{N}(\neg\varphi \to \mathsf{A}\neg\varphi)$. Thus, either $(d, a) \nVdash \varphi$ or $(d, a) \Vdash \mathsf{N}(\neg\varphi \to \mathsf{A}\neg\varphi)$. In the first case, $(d, a) \nVdash \mathsf{D}\varphi$ by item 6 above.

Next assume that $(d, a) \Vdash \mathsf{N}(\neg\varphi \to \mathsf{A}\neg\varphi)$. By item 6, to prove $(d, a) \nVdash \mathsf{D}\varphi$, it suffices to show that for any action $d' \in \mathcal{D}$ of the defender there is a response action $a' \in \mathcal{A}_{d'}$ of the attacker, such that $(d', a') \Vdash \varphi$. Indeed, consider any action $d' \in \mathcal{D}$ of the defender. By Definition 1, set $\mathcal{A}_{d'}$ is not empty. Let $a_1 \in \mathcal{A}_{d'}$ be an arbitrary response action of the attacker on action d'. Assumption $(d, a) \Vdash \mathsf{N}(\neg\varphi \to \mathsf{A}\neg\varphi)$, by item 4 of Definition 2, implies $(d', a_1) \Vdash \neg\varphi \to \mathsf{A}\neg\varphi$. We consider the following two cases separately:

Case I: $(d', a_1) \Vdash \varphi$. Then, choose the response action a' to be a_1 to have $(d', a') \Vdash \varphi$.

Case II: $(d', a_1) \nVdash \varphi$. Thus, $(d', a_1) \Vdash \neg\varphi$ by item 2 of Definition 2. Hence, $(d', a_1) \Vdash \mathsf{A}\neg\varphi$ by item 3 of Definition 2 because $(d', a_1) \Vdash \neg\varphi \to \mathsf{A}\neg\varphi$. Thus, by item 5 of Definition 2, there is a response action $a_2 \in \mathcal{A}_{d'}$ of the attacker such that $(d', a_2) \nVdash \neg\varphi$. Hence, $(d', a_2) \Vdash \varphi$ by item 2 of Definition 2. Then, choose the response action a' to be a_2 to have $(d', a') \Vdash \varphi$.

$(\Leftarrow)$: Suppose that $(d, a) \Vdash \varphi \wedge \neg\mathsf{N}(\neg\varphi \to \mathsf{A}\neg\varphi)$. Thus,

$$(d, a) \Vdash \varphi \tag{1}$$

and $(d, a) \nVdash \mathsf{N}(\neg\varphi \to \mathsf{A}\neg\varphi)$. The latter, by item 4 of Definition 2, implies that there is an action $d' \in \mathcal{D}$ of the defender and a response action $a' \in \mathcal{A}_{d'}$ of the attacker such that $(d', a') \nVdash \neg\varphi \to \mathsf{A}\neg\varphi$. Thus, $(d', a') \Vdash \neg\varphi$ and $(d', a') \nVdash \mathsf{A}\neg\varphi$ by item 3 of Definition 2. Then, $(d', a'') \Vdash \neg\varphi$ for each response action $a'' \in \mathcal{A}_{d'}$ of the attacker, by item 5 of Definition 2. Thus, $(d', a'') \nVdash \varphi$ for each response action $a'' \in \mathcal{A}_{d'}$ of the attacker, by item 2 of Definition 2. Hence, there exists an action $d' \in \mathcal{D}$ of the defender such that $(d', a'') \nVdash \varphi$ for each response action $a'' \in \mathcal{A}_{d'}$ of the attacker. Therefore, statement (1) implies $(d, a) \Vdash \mathsf{D}\varphi$ by item 6 above. ⊠

Axioms

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

1. Truth: $\Box\varphi \to \varphi$, where $\Box \in \{\mathsf{N}, \mathsf{A}\}$,
2. Negative Introspection: $\neg\mathsf{N}\varphi \to \mathsf{N}\neg\mathsf{N}\varphi$,
3. Distributivity: $\mathsf{N}(\varphi \to \psi) \to (\mathsf{N}\varphi \to \mathsf{N}\psi)$,
4. Unavoidability: $\mathsf{N}\varphi \to \neg\mathsf{A}\varphi$,
5. Strict Conditional: $\mathsf{N}(\varphi \to \psi) \to (\mathsf{A}\psi \to (\varphi \to \mathsf{A}\varphi))$,
6. Conjunction: $\mathsf{A}(\varphi \wedge \psi) \to (\mathsf{A}\varphi \vee \mathsf{A}\psi)$,
7. No Blame: $\neg\mathsf{A}(\varphi \to \mathsf{A}\varphi)$.

The Truth (for N), the Negative Introspection, and the Distributivity axioms are the well known S5 properties of the necessity modality N. The Truth axiom for modality A states that the attacker can only be blamed for something true. The Unavoidability axiom states that the attacker cannot be blamed for something that could not be prevented.

The Strict Conditional axiom states that if statement ψ is true under each action profile where φ is true, the attacker is blameable for ψ, and φ is true, then the attacker is also blamable for φ. Indeed, because statement ψ is true under each action profile where φ is true, any action of the attacker that prevents ψ also prevents φ. Hence, if the attacker is blameable for ψ and φ is true, then the attacker is also blamable for φ. We formalize this argument in Lemma 11.

The Truth axiom, the Unavoidability axiom, and the Strict Conditional axiom hold not only for modality A, but for modality D as well. These axioms are also true for strategic games.

The Conjunction and the No Blame axioms are the key axioms of our logical system. They capture the *asymmetry of* *information* in security games. Both of these axioms are true for the attacker's blameworthiness modality A – their soundness is proven in the appendix. However, as Lemma 3 and Lemma 4 show, they are not true for the defender's blameworthiness modality D in game G_2 depicted in Figure 4. Lemma 2 is an auxiliary statement about game G_2 used in the proofs of these two lemmas.

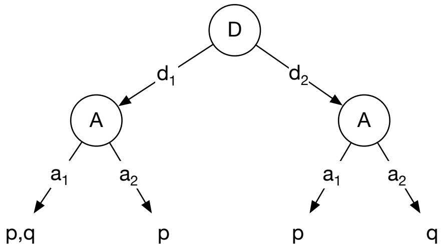

Figure 4: Game G_2, where $(d_1,a_1) \nVdash \mathsf{D}(p \wedge q) \to (\mathsf{D}p \vee \mathsf{D}q)$, $(d_2,a_2) \Vdash \mathsf{D}(p \to \mathsf{D}p)$, and $(d_2,a_1) \nVdash \mathsf{A}p \to \mathsf{N}(p \to \mathsf{A}p)$.

Lemma 2 *$(d,a) \nVdash \mathsf{D}p$ and $(d,a) \nVdash \mathsf{D}q$ for each action d of the defender and each response action a of the attacker in game G_2.*

PROOF. Note that $(d_1,a_1) \Vdash p$ and $(d_2,a_1) \Vdash p$, see Figure 4. Thus, for each action d' of the defender there is an action a' of the attacker such that $(d',a') \Vdash p$. Hence, $(d,a) \nVdash \mathsf{D}p$ by item 6 after Definition 2. Similarly, $(d_1,a_1) \Vdash q$ and $(d_2,a_2) \Vdash q$ imply that $(d,a) \nVdash \mathsf{D}q$. $\boxtimes$

Lemma 3 $(d_1,a_1) \nVdash \mathsf{D}(p \wedge q) \to (\mathsf{D}p \vee \mathsf{D}q)$.

PROOF. By Lemma 2, it suffices to show that $(d_1,a_1) \Vdash \mathsf{D}(p \wedge q)$. Indeed, observe that $(d_2,a_1) \nVdash p \wedge q$ and $(d_2,a_2) \nVdash p \wedge q$, see Figure 4. Thus, $(d_2,a) \nVdash p \wedge q$ for each response action a of the attacker on action d_2 of the defender. Also, $(d_1,a_1) \Vdash p \wedge q$, see Figure 4. Therefore, $(d_1,a_1) \Vdash \mathsf{D}(p \wedge q)$ by item 6 after Definition 2. $\boxtimes$

Lemma 4 $(d_2,a_2) \Vdash \mathsf{D}(p \to \mathsf{D}p)$.

PROOF. $(d_1,a_1) \nVdash \mathsf{D}p$ and $(d_1,a_2) \nVdash \mathsf{D}p$ by Lemma 2. Thus, $(d_1,a_1) \nVdash p \to \mathsf{D}p$ and $(d_1,a_2) \nVdash p \to \mathsf{D}p$ by item 3 of Definition 2 and because $(d_1,a_1) \Vdash p$ and $(d_1,a_2) \Vdash p$, see Figure 4. Thus, $(d_1,a) \nVdash p \to \mathsf{D}p$ for each response action a of the attacker on action d of the defender. At the same time, $(d_2,a_2) \Vdash p \to \mathsf{D}p$ by item 3 of Definition 2 because $(d_2,a_2) \nVdash p$, see Figure 4. Therefore, $(d_2,a_2) \Vdash \mathsf{D}(p \to \mathsf{D}p)$ by item 6 after Definition 2. $\boxtimes$

Informally, the Conjunction and the No Blame axioms capture the properties of the asymmetry of the information in security games and thus they cannot be true in strategic games (Naumov and Tao 2019) where the information is symmetric. A strategic game in which these axioms fail could be constructed by modifying the security game G_2 into a strategic game.

The logical system for blameworthiness in strategic games (Naumov and Tao 2019) includes the Fairness axiom: $\mathsf{B}_C\varphi \to \mathsf{N}(\varphi \to \mathsf{B}_C\varphi)$. In the next two lemmas we show that in the case of security games this axiom is not sound for modality A, but is sound for modality D.

Lemma 5 *$(d_2,a_1) \nVdash \mathsf{A}p \to \mathsf{N}(p \to \mathsf{A}p)$ in game G_2.*

PROOF. Note that $(d_2,a_1) \Vdash p$ and $(d_2,a_2) \nVdash p$, see Figure 4. Thus, $(d_2,a_1) \Vdash \mathsf{A}p$ by item 5 of Definition 2. Suppose that $(d_2,a_1) \Vdash \mathsf{A}p \to \mathsf{N}(p \to \mathsf{A}p)$. Hence, $(d_2,a_1) \Vdash \mathsf{N}(p \to \mathsf{A}p)$ by item 3 of Definition 2. Thus, $(d_1,a_1) \Vdash p \to \mathsf{A}p$ by item 4 of Definition 2. Note that $(d_1,a_1) \Vdash p$, see Figure 4. Hence, $(d_1,a_1) \Vdash \mathsf{A}p$ by item 3 of Definition 2. Then, by item 5 of Definition 2, there must exists a response action $a' \in \mathcal{D}_{d_1}$ of the attacker such that $(d_1,a') \nVdash p$. However, such an action a' does not exist because $(d_1,a_1) \Vdash p$ and $(d_1,a_2) \Vdash p$, see Figure 4. $\boxtimes$

Lemma 6 *$(d,a) \Vdash \mathsf{D}\varphi \to \mathsf{N}(\varphi \to \mathsf{D}\varphi)$ for any formula $\varphi \in \Phi^+$, any defender's action $d \in \mathcal{D}$, and any attacker's response action $a \in \mathcal{A}_d$ in an arbitrary security game $(\mathcal{D}, \{\mathcal{A}_d\}_{d \in \mathcal{D}}, \pi)$.*

PROOF. Suppose that $(d,a) \nVdash \mathsf{D}\varphi \to \mathsf{N}(\varphi \to \mathsf{D}\varphi)$. Thus, $(d,a) \Vdash \mathsf{D}\varphi$ and $(d,a) \nVdash \mathsf{N}(\varphi \to \mathsf{D}\varphi)$ by item 3 of Definition 2. By item 6 after Definition 2, statement $(d,a) \Vdash \mathsf{D}\varphi$, implies that $(d,a) \Vdash \varphi$.

By item 4 of Definition 2, statement $(d,a) \nVdash \mathsf{N}(\varphi \to \mathsf{D}\varphi)$ implies that there is an action $d_1 \in \mathcal{D}$ of the defender and a response action $a_1 \in \mathcal{A}_{d_1}$ of the attacker such that $(d_1,a_1) \nVdash \varphi \to \mathsf{D}\varphi$. Thus, $(d_1,a_1) \Vdash \varphi$ and $(d_1,a_1) \nVdash \mathsf{D}\varphi$ by item 3 of Definition 2. Hence, by item 6 after Definition 2, for each action $d' \in \mathcal{D}$ of the defender there is a response action $a' \in \mathcal{A}_{d'}$ of the attacker such that $(d',a') \Vdash \varphi$. Then, $(d,a) \nVdash \mathsf{D}\varphi$ by item 6 after Definition 2 because $(d,a) \Vdash \varphi$, which is a contradiction. $\boxtimes$

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

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

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

We conclude this section with an example of a formal proof in our logical system. The lemma below is used later in the proof of the completeness.

Lemma 7 *If $\vdash \varphi \leftrightarrow \psi$, then $\vdash \mathsf{A}\varphi \to \mathsf{A}\psi$.*

PROOF. By the Strict Conditional axiom,

$$\vdash \mathsf{N}(\psi \to \varphi) \to (\mathsf{A}\varphi \to (\psi \to \mathsf{A}\psi)).$$

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

$$\vdash \mathsf{A}\varphi \to (\psi \to \mathsf{A}\psi).$$

Thus, by the laws of propositional reasoning,

$$\vdash (\mathsf{A}\varphi \to \psi) \to (\mathsf{A}\varphi \to \mathsf{A}\psi). \tag{2}$$

Note that $\vdash \mathsf{A}\varphi \to \varphi$ by the Truth axiom. At the same time, $\vdash \varphi \leftrightarrow \psi$ by the assumption of the lemma. Thus, by the laws of propositional reasoning, $\vdash \mathsf{A}\varphi \to \psi$. Therefore, $\vdash \mathsf{A}\varphi \to \mathsf{A}\psi$ by the Modus Ponens inference rule from statement (2). ⊠

Soundness

In this section we prove the soundness of our logical system. The soundness of the Truth, the Negative Introspection, and the Distributivity axioms and of the two inference rules is straightforward. Below we prove the soundness of each of the remaining axioms as a separate lemma for any action $d \in \mathcal{D}$ of the defender, any response action $a \in \mathcal{A}_d$ of the attacker of an arbitrary security game $(\mathcal{D}, \{\mathcal{A}_d\}_{d\in\mathcal{D}}, \pi)$ and any formulae $\varphi, \psi \in \Phi$.

Lemma 8 *If* $(d,a) \Vdash \mathsf{N}\varphi$*, then* $(d,a) \nVdash \mathsf{A}\varphi$.

PROOF. By item 4 of Definition 2, the assumption $(d,a) \Vdash \mathsf{N}\varphi$ implies that $(d',a') \Vdash \varphi$ for each action $d' \in \mathcal{D}$ of the defender and each response action $a' \in \mathcal{A}_{d'}$ of the attacker. In particular, $(d,a') \Vdash \varphi$ for each response action $a' \in \mathcal{A}_d$ of the attacker. Therefore, $(d,a) \nVdash \mathsf{A}\varphi$ by item 5 of Definition 2. ⊠

Lemma 9 *If* $(d,a) \Vdash \mathsf{A}(\varphi \wedge \psi)$*, then either* $(d,a) \Vdash \mathsf{A}\varphi$ *or* $(d,a) \Vdash \mathsf{A}\psi$.

PROOF. By item 5 of Definition 2, the assumption $(d,a) \Vdash \mathsf{A}(\varphi \wedge \psi)$ implies that $(d,a) \Vdash \varphi \wedge \psi$ and there is a response action $a' \in \mathcal{A}_d$ of the attacker such that $(d,a') \nVdash \varphi \wedge \psi$. Hence, either $(d,a') \nVdash \varphi$ or $(d,a') \nVdash \psi$. Without loss of generality, suppose that $(d,a') \nVdash \varphi$. At the same time, statement $(d,a) \Vdash \varphi \wedge \psi$ implies that $(d,a) \Vdash \varphi$. Hence, $(d,a) \Vdash \varphi$ and $(d,a') \nVdash \varphi$. Therefore, $(d,a) \Vdash \mathsf{A}\varphi$ by item 5 of Definition 2. ⊠

Lemma 10 $(d,a) \nVdash \mathsf{A}(\varphi \to \mathsf{A}\varphi)$.

PROOF. Suppose that $(d,a) \Vdash \mathsf{A}(\varphi \to \mathsf{A}\varphi)$. Thus, by item 5 of Definition 2,

$$(d,a) \Vdash \varphi \to \mathsf{A}\varphi \tag{3}$$

and there is a response action $a' \in \mathcal{A}_d$ of the attacker such that $(d,a') \nVdash \varphi \to \mathsf{A}\varphi$. Hence, $(d,a') \Vdash \varphi$ and $(d,a') \nVdash \mathsf{A}\varphi$ by item 3 of Definition 2. Thus,

$$(d,a'') \Vdash \varphi \tag{4}$$

for any response action $a'' \in \mathcal{A}_d$ of the attacker, by item 5 of Definition 2. In particular, $(d,a) \Vdash \varphi$. Then, $(d,a) \Vdash \mathsf{A}\varphi$ due to statement (3) and item 3 of Definition 2. Thus, by item 5 of Definition 2, there must exist a response action $b \in \mathcal{A}_d$ of the attacker such that $(d,b) \nVdash \varphi$, which contradicts to statement (4). ⊠

Lemma 11 *If* $(d,a) \Vdash \mathsf{N}(\varphi \to \psi)$*,* $(d,a) \Vdash \mathsf{A}\psi$*, and* $(d,a) \Vdash \varphi$*, then* $(d,a) \Vdash \mathsf{A}\varphi$.

PROOF. By item 5 of Definition 2, the assumption $(d,a) \Vdash \mathsf{A}\psi$ implies that there is a response action $a' \in \mathcal{A}_d$ of the attacker such that $(d,a') \nVdash \psi$. At the same time, $(d,a') \Vdash \varphi \to \psi$ by item 4 of Definition 2 and the assumption $(d,a) \Vdash \mathsf{N}(\varphi \to \psi)$. Hence, $(d,a') \nVdash \varphi$ by item 3 of Definition 2. Therefore, $(d,a) \Vdash \mathsf{A}\varphi$ by the assumption $(d,a) \Vdash \varphi$ and item 5 of Definition 2. ⊠

Completeness

In this section we prove the completeness of our logical system in three steps. First, we introduce an auxiliary modality R as an abbreviation definable through modality A. Next, we define a canonical security game and prove its basic property. Finally, we state and prove the strong completeness theorem for our logical system.

Preliminaries

Let $\mathsf{R}\varphi$ be an abbreviation for $\neg(\varphi \to \mathsf{A}\varphi)$. Note that $\mathsf{R}\varphi$ stands for "statement φ is true, but the attacker cannot be blamed for it". In other words, $\mathsf{R}\varphi$ means that *the defender's* action *unavoidably* led to φ being true. This modality is not present in (Naumov and Tao 2019). In the context of STIT logic, but not in the context of security games, a similar single-agent modality was studied in (Xu 1998). The same modality for coalitions was investigated in (Broersen, Herzig, and Troquard 2009). Below we prove the key properties of modality R that are used later in the proof of the completeness.

Lemma 12 $\vdash \mathsf{N}\varphi \to \mathsf{R}\varphi$.

PROOF. By the Unavoidability axiom, $\vdash \mathsf{N}\varphi \to \neg\mathsf{A}\varphi$. At the same time, $\vdash \mathsf{N}\varphi \to \varphi$ by the Truth axiom. Hence, by propositional reasoning, $\vdash \mathsf{N}\varphi \to \varphi \wedge \neg\mathsf{A}\varphi$. Thus, again by propositional reasoning, $\vdash \mathsf{N}\varphi \to \neg(\varphi \to \mathsf{A}\varphi)$. Therefore, $\vdash \mathsf{N}\varphi \to \mathsf{R}\varphi$ by the definition of modality R. ⊠

The next four lemmas show that R is an S5 modality.

Lemma 13 *Inference rule* $\dfrac{\varphi}{\mathsf{R}\varphi}$ *is derivable.*

PROOF. Suppose that $\vdash \varphi$. Thus, $\vdash \mathsf{N}\varphi$ by the Necessitation inference rule. Therefore, $\vdash \mathsf{R}\varphi$ by Lemma 12 and the Modus Ponens inference rule. ⊠

Lemma 14 $\vdash \mathsf{R}\varphi \to \varphi$.

PROOF. Note that formula $\neg(\varphi \to \mathsf{A}\varphi) \to \varphi$ is a propositional tautology. Thus, $\vdash \mathsf{R}\varphi \to \varphi$ by the definition of the modality R. ⊠

Lemma 15 $\vdash \mathsf{R}(\varphi \to \psi) \to (\mathsf{R}\varphi \to \mathsf{R}\psi)$.

PROOF. Note that the following formula is a propositional tautology

$$\neg((\varphi \to \psi) \to \mathsf{A}(\varphi \to \psi)) \to \\ (\neg(\varphi \to \mathsf{A}\varphi) \to (\neg\mathsf{A}(\varphi \to \psi) \wedge \neg\mathsf{A}\varphi)).$$

Thus, it follows from the definition of the modality R that

$$\vdash \mathsf{R}(\varphi\to\psi)\to(\mathsf{R}\varphi\to(\neg\mathsf{A}(\varphi\to\psi)\wedge\neg\mathsf{A}\varphi)).$$

At the same time, formula

$$(\neg\mathsf{A}(\varphi\to\psi)\wedge\neg\mathsf{A}\varphi)\to\neg\mathsf{A}((\varphi\to\psi)\wedge\varphi)$$

is a contrapositive of the Conjunction axiom. Thus, by the laws of propositional reasoning,

$$\vdash \mathsf{R}(\varphi\to\psi)\to(\mathsf{R}\varphi\to\neg\mathsf{A}((\varphi\to\psi)\wedge\varphi).\tag{5}$$

Next, note that the following formula is also a propositional tautology $((\varphi\to\psi)\wedge\varphi)\to\psi$. Hence, by the Necessitation inference rule, $\vdash \mathsf{N}(((\varphi\to\psi)\wedge\varphi)\to\psi)$. Thus, by the Strict Conditional axiom and the Modus Ponens inference rule,

$$\vdash \mathsf{A}\psi\to((\varphi\to\psi)\wedge\varphi\to\mathsf{A}((\varphi\to\psi)\wedge\varphi)).$$

Then, by the laws of propositional reasoning,

$$\vdash\neg\mathsf{A}((\varphi\to\psi)\wedge\varphi)\to((\varphi\to\psi)\wedge\varphi\to\neg\mathsf{A}\psi).$$

Hence, by propositional reasoning using statement (5),

$$\vdash \mathsf{R}(\varphi\to\psi)\to(\mathsf{R}\varphi\to((\varphi\to\psi)\wedge\varphi\to\neg\mathsf{A}\psi)).\tag{6}$$

Note that the following formula is a propositional tautology

$$\neg((\varphi\to\psi)\to\mathsf{A}(\varphi\to\psi))\to$$
$$(\neg(\varphi\to\mathsf{A}\varphi)\to((\varphi\to\psi)\wedge\varphi)).$$

Thus, it follows from the definition of the modality R that

$$\vdash \mathsf{R}(\varphi\to\psi)\to(\mathsf{R}\varphi\to((\varphi\to\psi)\wedge\varphi)).\tag{7}$$

Then, by propositional reasoning using statement (6),

$$\vdash \mathsf{R}(\varphi\to\psi)\to(\mathsf{R}\varphi\to\neg\mathsf{A}\psi).\tag{8}$$

Additionally, note that $((\varphi\to\psi)\wedge\varphi)\to\psi$ is a propositional tautology. Hence, statement (7) also implies

$$\vdash \mathsf{R}(\varphi\to\psi)\to(\mathsf{R}\varphi\to\psi).$$

Thus, by propositional reasoning using statement (8),

$$\vdash \mathsf{R}(\varphi\to\psi)\to(\mathsf{R}\varphi\to(\psi\wedge\neg\mathsf{A}\psi)).$$

Again by propositional reasoning,

$$\vdash \mathsf{R}(\varphi\to\psi)\to(\mathsf{R}\varphi\to\neg(\psi\to\mathsf{A}\psi)).$$

Therefore, $\vdash \mathsf{R}(\varphi\to\psi)\to(\mathsf{R}\varphi\to\mathsf{R}\psi)$ by the definition of the modality R. $\boxtimes$

Lemma 16 $\vdash\neg\mathsf{R}\varphi\to\mathsf{R}\neg\mathsf{R}\varphi$.

PROOF. Note that $\neg\neg(\varphi\to\mathsf{A}\varphi)\leftrightarrow(\varphi\to\mathsf{A}\varphi)$ is a propositional tautology. Thus, $\vdash\mathsf{A}\neg\neg(\varphi\to\mathsf{A}\varphi)\to\mathsf{A}(\varphi\to\mathsf{A}\varphi)$ by Lemma 7. Hence, $\vdash\neg\mathsf{A}(\varphi\to\mathsf{A}\varphi)\to\neg\mathsf{A}\neg\neg(\varphi\to\mathsf{A}\varphi)$ by contraposition. Then, $\vdash\neg\mathsf{A}\neg\neg(\varphi\to\mathsf{A}\varphi)$ by the No Blame Axiom and the Modus Ponens inference rule. Thus, by the laws of propositional reasoning,

$$\vdash(\varphi\to\mathsf{A}\varphi)\to\neg((\varphi\to\mathsf{A}\varphi)\to\mathsf{A}\neg\neg(\varphi\to\mathsf{A}\varphi)).$$

Hence, again by the laws of propositional reasoning,

$$\vdash\neg\neg(\varphi\to\mathsf{A}\varphi)\to\neg(\neg\neg(\varphi\to\mathsf{A}\varphi)\to\mathsf{A}\neg\neg(\varphi\to\mathsf{A}\varphi)).$$

Recall that $\mathsf{R}\varphi$ is an abbreviation for $\neg(\varphi\to\mathsf{A}\varphi)$. Then,

$$\vdash\neg\mathsf{R}\varphi\to\neg(\neg\mathsf{R}\varphi\to\mathsf{A}\neg\mathsf{R}\varphi).$$

Thus, $\vdash\neg\mathsf{R}\varphi\to\mathsf{R}\neg\mathsf{R}\varphi$ again by the definition of R. $\boxtimes$

The next two lemmas capture well known properties of S5 modalities.

Lemma 17 *If $\varphi_1,\dots,\varphi_n\vdash\psi$, then $\Box\varphi_1,\dots,\Box\varphi_n\vdash\Box\psi$, where $\Box$ is either modality N or modality R.*

PROOF. First, consider the case when $\Box$ is modality N. Assumption $\varphi_1,\dots,\varphi_n\vdash\psi$ by the deduction lemma implies that $\vdash\varphi_1\to(\varphi_2\to\dots(\varphi_n\to\psi)\dots)$. Hence, by the Necessitation rule, $\vdash\mathsf{N}(\varphi_1\to(\varphi_2\to\dots(\varphi_n\to\psi)\dots))$. Thus, by the Distributivity axiom and the Modus Ponens inference rule, $\vdash\mathsf{N}\varphi_1\to\mathsf{N}(\varphi_2\to\dots(\varphi_n\to\psi)\dots)$. Hence, $\mathsf{N}\varphi_1\vdash\mathsf{N}(\varphi_2\to\dots(\varphi_n\to\psi)\dots)$ again by the Modus Ponens inference rule. By repeating the previous two steps $(n-1)$ more times, $\mathsf{N}\varphi_1,\dots,\mathsf{N}\varphi_n\vdash\mathsf{N}\psi$.

The case when $\Box$ is modality R is similar, but it uses Lemma 13 instead of the Necessitation inference rule and Lemma 15 instead of the Distributivity axiom. $\boxtimes$

Lemma 18 $\vdash\Box\varphi\to\Box\Box\varphi$ *where $\Box$ is either modality N or modality R.*

PROOF. We first consider the case when $\Box$ is modality N. Formula $\mathsf{N}\neg\mathsf{N}\varphi\to\neg\mathsf{N}\varphi$ is an instance of the Truth axiom. Thus, $\vdash\mathsf{N}\varphi\to\neg\mathsf{N}\neg\mathsf{N}\varphi$ by contraposition. Hence, taking into account the following instance of the Negative Introspection axiom: $\neg\mathsf{N}\neg\mathsf{N}\varphi\to\mathsf{N}\neg\mathsf{N}\neg\mathsf{N}\varphi$, we have

$$\vdash\mathsf{N}\varphi\to\mathsf{N}\neg\mathsf{N}\neg\mathsf{N}\varphi.\tag{9}$$

At the same time, $\neg\mathsf{N}\varphi\to\mathsf{N}\neg\mathsf{N}\varphi$ is an instance of the Negative Introspection axiom. Thus, $\vdash\neg\mathsf{N}\neg\mathsf{N}\varphi\to\mathsf{N}\varphi$ by the law of contrapositive in the propositional logic. Hence, by the Necessitation inference rule, $\vdash\mathsf{N}(\neg\mathsf{N}\neg\mathsf{N}\varphi\to\mathsf{N}\varphi)$. Thus, by the Distributivity axiom and the Modus Ponens inference rule, $\vdash\mathsf{N}\neg\mathsf{N}\neg\mathsf{N}\varphi\to\mathsf{N}\mathsf{N}\varphi$. The latter, together with statement (9), implies the statement of the lemma by propositional reasoning.

The case when $\Box$ is modality R is similar, but it uses Lemma 14 instead of the Truth axiom, Lemma 16 instead of the Negative Introspection axiom, Lemma 13 instead of the Necessitation inference rule, and Lemma 15 instead of the Distributivity axiom. $\boxtimes$

Canonical Security Game

We define the canonical game $G(X)=(\Omega,\{\mathcal{A}_\delta\}_{\delta\in\Omega},\pi)$ for each maximal consistent set of formulae X.

Definition 3 *Ω is the set of all maximal consistent sets of formulae such that if $\omega\in\Omega$, then $\{\varphi\in\Phi\mid\mathsf{N}\varphi\in X\}\subseteq\omega$.*

Definition 4 *$\omega\sim\omega'$ if $\forall\varphi\in\Phi\ (\mathsf{R}\varphi\in\omega\Leftrightarrow\mathsf{R}\varphi\in\omega')$.*

Note that $\sim$ is an equivalence relation on set Ω. The set $\mathcal{A}_\delta$ of possible responses by the attacker on an action $\delta\in\Omega$ of the defender is the (nonempty) equivalence class of element δ with respect to this equivalence relation:

Definition 5 $\mathcal{A}_\delta=[\delta]$.

Thus, each defender's action $\delta\in\Omega$ and each attacker's responses $\omega\in[\delta]$ are maximal consistent sets of formulae. This is significantly different from (Naumov and Tao 2019), where actions of all agents are formulae.

Definition 6 $\pi(p)=\{(\delta,\omega)\in\Omega\times\Omega\mid\omega\in\mathcal{A}_\delta,p\in\omega\}$.

This concludes the definition of the canonical game $G(X)$.

As usual, at the core of the proof of completeness is a truth lemma (or an induction lemma), which in our case is Lemma 23. The next four lemmas are auxiliary statements used in the induction step of the proof of Lemma 23.

Lemma 19 *For any action $\delta \in \Omega$ of the defender, any response action $\omega \in [\delta]$ of the attacker, and any formula $\mathsf{A}\varphi \in \omega$, we have (i) $\varphi \in \omega$ and (ii) there is a response action $\omega' \in [\delta]$ such that $\varphi \notin \omega'$.*

PROOF. Assumption $\mathsf{A}\varphi \in \omega$ implies that $\omega \vdash \varphi$ by the Truth axiom and the Modus Ponens inference rule. Thus, $\varphi \in \omega$ because set ω is maximal. This concludes the proof of the first statement. To prove the second statement, consider the set of formulae

$$Y = \{\neg\varphi\} \cup \{\psi \mid \mathsf{R}\psi \in \omega\} \cup \{\chi \mid \mathsf{N}\chi \in \omega\}. \tag{10}$$

Claim 1 *Set Y is consistent.*

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

$$\mathsf{R}\psi_1, \dots, \mathsf{R}\psi_k, \mathsf{N}\chi_1, \dots, \mathsf{N}\chi_n \in \omega \tag{11}$$

such that $\psi_1, \dots, \psi_k, \chi_1, \dots, \chi_n \vdash \varphi$. Hence, by Lemma 17, $\mathsf{R}\psi_1, \dots, \mathsf{R}\psi_k, \mathsf{R}\chi_1, \dots, \mathsf{R}\chi_n \vdash \mathsf{R}\varphi$. Then, by Lemma 12 and the Modus Ponens inference rule, $\mathsf{R}\psi_1, \dots, \mathsf{R}\psi_k, \mathsf{N}\chi_1, \dots, \mathsf{N}\chi_n \vdash \mathsf{R}\varphi$. Thus, $\omega \vdash \mathsf{R}\varphi$ by statement (11). Hence, $\omega \vdash \neg(\varphi \to \mathsf{A}\varphi)$ by the definition of the modality R. Then, $\omega \vdash \neg\mathsf{A}\varphi$ by the laws of the propositional reasoning, which contradicts the assumption $\mathsf{A}\varphi \in \omega$ of the lemma because set ω is consistent. $\boxtimes$

Let set ω' be any maximal consistent extension of set Y. Then, $\neg\varphi \in \omega'$. Thus, $\varphi \notin \omega'$ because set ω' is consistent.

Claim 2 $\omega' \in \Omega$.

PROOF OF CLAIM. Consider any formula $\mathsf{N}\chi \in X$. By Definition 3, it suffices to show that $\chi \in \omega'$. Indeed, assumption $\mathsf{N}\chi \in X$ implies that $X \vdash \mathsf{N}\mathsf{N}\chi$ by Lemma 18. Thus, $\mathsf{N}\mathsf{N}\chi \in X$ because set X is maximal. Then, $\mathsf{N}\chi \in \omega$ by Definition 3 and the assumption $\omega \in [\delta] \subseteq \Omega$ of the lemma. Hence, $\chi \in Y \subseteq \omega'$ by equation (10) and the choice of set ω'. $\boxtimes$

Claim 3 $\omega' \in [\delta]$.

PROOF OF CLAIM. Recall that $\omega \in [\delta]$ by the assumption of the lemma. Thus, by Claim 2, it suffices to show that $\omega \sim \omega'$. Hence, by Definition 4, it suffices to prove that $\mathsf{R}\psi \in \omega$ iff $\mathsf{R}\psi \in \omega'$ for each formula $\psi \in \Phi$. If $\mathsf{R}\psi \in \omega$, then $\omega \vdash \mathsf{R}\mathsf{R}\psi$ by Lemma 18. Hence, $\mathsf{R}\mathsf{R}\psi \in \omega$ because set ω is maximal. Thus, $\mathsf{R}\psi \in Y \subseteq \omega'$ by equation (10) and the choice of ω'.

Suppose that $\mathsf{R}\psi \notin \omega$. Thus, $\omega \vdash \mathsf{R}\neg\mathsf{R}\psi$ by Lemma 16 and the Modus Ponens inference rule. Hence, $\mathsf{R}\neg\mathsf{R}\psi \in \omega$ because set ω is maximal. Thus, $\neg\mathsf{R}\psi \in Y \subseteq \omega'$ by equation (10) and the choice of set ω'. Therefore, $\mathsf{R}\psi \notin \omega'$ because set ω' is consistent. $\boxtimes$

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

Lemma 20 *For any action $\delta \in \Omega$ of the defender, any response action $\omega \in [\delta]$ of the attacker, and any formula $\varphi \in \Phi$, if $\neg(\varphi \to \mathsf{A}\varphi) \in \omega$, then $\varphi \in \omega'$ for each $\omega' \in [\delta]$.*

PROOF. Assumption $\neg(\varphi \to \mathsf{A}\varphi) \in \omega$ implies $\mathsf{R}\varphi \in \omega$ by the definition of the modality R. Note that $\omega \sim \omega'$ because $\omega, \omega' \in [\delta]$. Thus, $\mathsf{R}\varphi \in \omega'$ by Definition 4. Hence, $\omega' \vdash \varphi$ by Lemma 14 and the Modus Ponens inference rule. Therefore, $\varphi \in \omega'$ because set ω' is maximal. $\boxtimes$

Lemma 21 *For any actions $\omega, \omega' \in \Omega$, if $\mathsf{N}\varphi \in \omega$, then $\varphi \in \omega'$.*

PROOF. Suppose that $\varphi \notin \omega'$. Hence, $\mathsf{N}\varphi \notin X$ by Definition 3 and the assumption $\omega' \in \Omega$. Thus, $\neg\mathsf{N}\varphi \in X$ because set X is maximal. Then, $X \vdash \mathsf{N}\neg\mathsf{N}\varphi$ by the Negative Introspection axiom and the Modus Ponens inference rule. Hence, $\mathsf{N}\neg\mathsf{N}\varphi \in X$ again because set X is maximal. Thus, $\neg\mathsf{N}\varphi \in \omega$ by Definition 3 and the assumption $\omega \in \Omega$. Therefore, $\mathsf{N}\varphi \notin \omega$ because set ω is consistent. $\boxtimes$

Lemma 22 *For any action $\omega \in \Omega$ and any formula $\neg\mathsf{N}\varphi \in \omega$, there is an action $\omega' \in \Omega$ such that $\varphi \notin \omega'$.*

PROOF. Consider the set of formulae

$$Y = \{\neg\varphi\} \cup \{\psi \mid \mathsf{N}\psi \in \omega\}. \tag{12}$$

Claim 4 *Set Y is consistent.*

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

$$\mathsf{N}\psi_1, \dots, \mathsf{N}\psi_n \in \omega \tag{13}$$

such that $\psi_1, \dots, \psi_n \vdash \varphi$. Hence, $\mathsf{N}\psi_1, \dots, \mathsf{N}\psi_n \vdash \mathsf{N}\varphi$ by Lemma 17. Thus, $\omega \vdash \mathsf{N}\varphi$ by the assumption (13), which contradicts the assumption $\neg\mathsf{N}\varphi \in \omega$ of the lemma because set ω is consistent. $\boxtimes$

Let set ω' be any maximal consistent extension of set Y. Then, $\neg\varphi \in \omega'$. Thus, $\varphi \notin \omega'$ because set ω' is consistent.

Claim 5 $\omega' \in \Omega$.

PROOF OF CLAIM. Consider any formula $\mathsf{N}\psi \in X$. By Definition 3, it suffices to show that $\psi \in \omega'$. Indeed, assumption $\mathsf{N}\psi \in X$ implies that $X \vdash \mathsf{N}\mathsf{N}\psi$ by Lemma 18. Thus, $\mathsf{N}\mathsf{N}\psi \in X$ because set X is maximal. Then, $\mathsf{N}\psi \in \omega$ by Definition 3 and the assumption $\omega \in \Omega$ of the lemma. Therefore, $\psi \in Y \subseteq \omega'$ by equation (12) and the choice of set ω'. $\boxtimes$

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

Lemma 23 (truth lemma) *For each formula φ, each action of the defender $\delta \in \Omega$, and each response action $\omega \in [\delta]$ of the attacker, $(\delta, \omega) \Vdash \varphi$ iff $\varphi \in \omega$.*

PROOF. We prove the lemma by structural induction on formula φ. The case when formula φ is a propositional variable follows from Definition 2 and Definition 6. The cases when formula φ is a negation or an implication follow from Definition 2 and the assumption of the maximality and the consistency of set ω in the standard way.

Suppose that formula φ has the form $\mathsf{A}\psi$.

$(\Rightarrow)$: Assume that $\mathsf{A}\psi \notin \omega$. Hence, $\omega \nvdash \mathsf{A}\psi$ because set ω is maximal. We consider the following two cases separately:

Case I: $(\psi \to \mathsf{A}\psi) \in \omega$. Thus, statement $\omega \nvdash \mathsf{A}\psi$ implies $\omega \nvdash \psi$ by the contraposition of the Modus Ponens inference

rule. Hence, $\psi \notin \omega$. Then, $(\delta, \omega) \nVdash \varphi$ by the induction hypothesis. Therefore, $(\delta, \omega) \nVdash \mathsf{A}\varphi$ by item 5 of Definition 2.
Case II: $(\psi \to \mathsf{A}\psi) \notin \omega$. Hence, $\neg(\psi \to \mathsf{A}\psi) \in \omega$ because set ω is maximal. Thus, $\psi \in \omega'$ for each action $\omega' \in [\delta]$, by Lemma 20. Then, by the induction hypothesis, $(\delta, \omega') \Vdash \psi$ for each response action $\omega' \in [\delta]$ of the attacker on action $\delta \in \Omega$ of the defender. Therefore, $(\delta, \omega) \nVdash \mathsf{A}\psi$ by item 5 of Definition 2.
$(\Leftarrow)$: Assume that $\mathsf{A}\psi \in \omega$. Thus, by Lemma 19, we have (i) $\psi \in \omega$ and (ii) there is a response action $\omega' \in [\delta]$ such that $\psi \notin \omega'$. Hence, by the induction hypothesis, (i) $(\delta, \omega) \Vdash \psi$ and (ii) there is a response action $\omega' \in [\delta]$ of the attacker such that $(\delta, \omega') \nVdash \psi$. Therefore, $(\delta, \omega) \Vdash \mathsf{A}\psi$ by item 5 of Definition 2.

Next, assume formula φ has the form $\mathsf{N}\psi$.
$(\Rightarrow)$: Let $\mathsf{N}\psi \notin \omega$. Thus, $\neg\mathsf{N}\psi \in \omega$ because set ω is maximal. Hence, by Lemma 22, there is an action $\omega' \in \Omega$ such that $\psi \notin \omega'$. Note that $\omega' \in [\omega']$ because $[\omega']$ is an equivalence class. Thus, $(\omega', \omega') \nVdash \psi$ by the induction hypothesis. Therefore, $(\delta, \omega) \nVdash \mathsf{N}\psi$ by item 4 of Definition 2.
$(\Leftarrow)$: Suppose that $\mathsf{N}\psi \in \omega$. Thus, $\psi \in \omega'$ for each action $\omega' \in \Omega$ by Lemma 21. Hence, by the induction hypothesis, $(\delta', \omega') \Vdash \psi$ for each action $\delta' \in \Omega$ of the defender and each response action $\omega' \in [\delta']$ of the attacker. Therefore, $(\delta, \omega) \Vdash \mathsf{N}\psi$ by item 4 of Definition 2. ⊠

Recall that the canonical game $G(X)$ is defined for an arbitrary maximal consistent set of formulae X.

Lemma 24 $X \in \Omega$.

PROOF. Consider any formula $\mathsf{N}\varphi \in X$. By Definition 3, it suffices to show that $\varphi \in X$. Indeed, assumption $\mathsf{N}\varphi \in X$ implies $X \vdash \varphi$ by the Truth axiom and the Modus Ponens inference rule. Thus, $\varphi \in X$ because set X is maximal. ⊠

Strong Completeness Theorem

Theorem 1 *If $X_0 \nvdash \varphi$, then there is an action $d \in \mathcal{D}$ of the defender and a response action $a \in \mathcal{A}_d$ of the attacker in a game $(\mathcal{D}, \{\mathcal{A}_d\}_{d\in\mathcal{D}}, \pi)$ such that $(d, a) \Vdash \chi$ for each formula $\chi \in X_0$ and $(d, a) \nVdash \varphi$.*

PROOF. Let the set of formulae $X \subseteq \Phi$ be any maximal consistent extension of set $X_0 \cup \{\neg\varphi\}$. Then, $\varphi \notin X$ because set X is consistent.

Consider the canonical game $G(X) = (\Omega, \{\mathcal{A}_\delta\}_{\delta\in\Omega}, \pi)$. Then, $X \in \Omega$ by Lemma 24. Also, $X \in [X] = \mathcal{A}_X$ because set $[X]$ is an equivalence class and because of Definition 5. Therefore, $(X, X) \Vdash \chi$ for each formula $\chi \in X_0 \subseteq X$ and $(X, X) \nVdash \varphi$ by Lemma 23. ⊠

Conclusion

In this paper we gave a sound and complete axiomatic system that describes the properties of blameworthiness in security games. A natural next step is to generalize this work to arbitrary extensive form games. The Conjunction and the No Blame axioms in this paper are specific to security games and are not sound for arbitrary extensive form games. As we have seen in Lemma 3 and Lemma 4, these axioms are already not sound for the player who makes the first move in a security game. Although these axioms are sound for the player making the second move in security games, it is not sound for the second player in an arbitrary extensive form game. Consider, for example, game G_3 depicted in Figure 5. In this game, $(d_1, a_2) \Vdash \mathsf{A}(p \wedge q)$ because formula $p \wedge q$ is true under the action profile (d_1, a_2), but the second player could have prevented it by using action a_1 instead of a_2. At the same time, $(d_1, a_2) \nVdash \mathsf{A}p \vee \mathsf{A}q$ because the second player has neither a strategy that would prevent p nor a strategy that would prevent q. This is a counterexample for the Conjunction axiom. The game G_3 also provides a counterexample for the No Blame axiom: $(d_1, a_1, d_1) \Vdash \mathsf{A}(p \to \mathsf{A}p)$. Indeed, $(d_1, a_1, d_1) \Vdash p \to \mathsf{A}p$ because $(d_1, a_1, d_1) \nVdash p$. At the same time, $(d_1, a_2) \nVdash p \to \mathsf{A}p$. Thus, the second player could have prevented $p \to \mathsf{A}p$ by using a_2 instead of a_1.

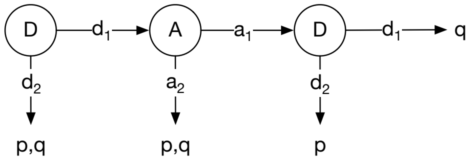

Figure 5: Game G_3, where $(d_1, a_2) \nVdash \mathsf{A}(p \wedge q) \to (\mathsf{A}p \vee \mathsf{A}q)$, and $(d_1, a_1, d_1) \Vdash \mathsf{A}(p \to \mathsf{A}p)$.

In addition to finding the right set of axioms, proving a completeness theorem would also require to recover the structure of the canonical game tree from a maximal consistent set of formulae. Finding the right set of axioms sound for all extensive form games and proving their completeness remains an open problem.

References

An, B.; Tambe, M.; and Sinha, A. 2016. Stackelberg security games (SSG) basics and application overview. In *Improving Homeland Security Decisions*. Cambridge Univ. Press.

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

Broersen, J.; Herzig, A.; and Troquard, N. 2009. What groups do, can do, and know they can do: an analysis in normal modal logics. *Journal of Applied Non-Classical Logics* 19(3):261–289.

Brown, M.; Sinha, A.; Schlenker, A.; and Tambe, M. 2016. One size does not fit all: A game-theoretic approach for dynamically and effectively screening for threats. In *Thirtieth AAAI Conference on Artificial Intelligence*.

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

Fang, F.; Stone, P.; and Tambe, M. 2015. When security games go green: Designing defender strategies to prevent poaching and illegal fishing. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

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

Halpern, J. Y. 2016. *Actual causality*. MIT Press.

Jain, M.; Tsai, J.; Pita, J.; Kiekintveld, C.; Rathi, S.; Tambe, M.; and Ordónez, F. 2010. Software assistants for randomized patrol planning for the LAX airport police and the federal air marshal service. *Interfaces* 40(4):267–290.

Lewis, D. 2013. *Counterfactuals*. John Wiley & Sons.

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

Pita, J.; Jain, M.; Marecki, J.; Ordóñez, F.; Portway, C.; Tambe, M.; Western, C.; Paruchuri, P.; and Kraus, S. 2008. Deployed ARMOR protection: the application of a game theoretic model for security at the Los Angeles International Airport. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems: industrial track*, 125–132. International Foundation for Autonomous Agents and Multiagent Systems.

Sinha, A.; Fang, F.; An, B.; Kiekintveld, C.; and Tambe, M. 2018. Stackelberg security games: Looking beyond a decade of success. In *IJCAI*, 5494–5501.

von Stackelberg, H. 1934. *Marktform und gleichgewicht*. J. Springer.

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

Xu, M. 1998. Axioms for deliberative stit. *Journal of Philosophical Logic* 27(5):505–552.