Responsibility in Extensive Form Games

Qi Shi[*]

University of Southampton
qi.shi@soton.ac.uk

Abstract

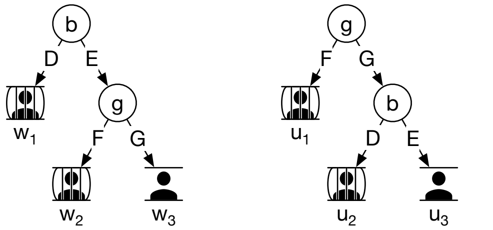
Two different forms of responsibility, counterfactual and seeing-to-it, have been extensively discussed in philosophy and AI in the context of a single agent or multiple agents acting simultaneously. Although the generalisation of counterfactual responsibility to a setting where multiple agents act in some order is relatively straightforward, the same cannot be said about seeing-to-it responsibility. Two versions of seeing-to-it modality applicable to such settings have been proposed in the literature. Neither of them perfectly captures the intuition of responsibility. This paper proposes a definition of seeing-to-it responsibility for such settings that amalgamate the two modalities.

This paper shows that the newly proposed notion of responsibility and counterfactual responsibility are not definable through each other and studies the responsibility gap for these two forms of responsibility. It shows that although these two forms of responsibility are not enough to ascribe responsibility in each possible situation, this gap does not exist if higher-order responsibility is taken into account.

1 Introduction

In the United States, if a person is found guilty by a state court and all appeals within the state justice system have been exhausted, the person can petition the Governor of the state for executive clemency. The US Supreme Court once described the clemency by the executive branch of the government as the “fail safe” of the criminal justice system (US Supreme Court 1993). This was the case with Barry Beach, who was found guilty of killing a 17-year-old high school valedictorian Kim Nees and sentenced in 1984 to 100 years imprisonment without parole (Associated Press 2015). In 2014, after a court appeal, a retrial, and a negative decision by the Montana Supreme Court, Barry’s attorney filed a petition for executive clemency.

To prevent corruption and favouritism by the Governor, many states in the US have boards that must support the decision before the Governor can grant executive clemency. In Montana, such a board has existed since the original 1889 Constitution (Constitution Convention 1889, Article VII, Section 9). With time, the law, the name of the board, and the way it grants approval changed (Montana Board of Pardons and Parole 2023), but the Board maintained the ability to constrain the Governor’s power to grant executive clemency. The executive clemency procedure that existed in Montana by 2014 is captured by the *extensive form game* depicted in Figure 1a. First, the Board (agent b) can either deny (action D) the clemency or recommend (E) it. If the Board recommends, then the Governor (agent g) might grant (G) or not grant (F) the executive clemency.



(a) Montana situation (b) Switched action order

Figure 1: Executive clemency procedure

The executive clemency procedure in Montana is a typical multiagent system where the final outcome is determined by the decisions of all agents in the system. Such multiagent systems widely exist in both human and machine activities and have been studied from multiple perspectives. *Responsibility* is one of the topics in those studies. Although there is no commonly acknowledged definition of responsibility, we usually have some vague intuition about the responsibility of an agent when we think of the agent being praiseworthy for a positive result or blameworthy for a negative result, especially if the agent is undertaking moral or legal obligations. It is also usually assumed that responsibility is connected to *free will* of an agent to act. Note that, an agent can act to prevent or to achieve a certain result. This gives rise to two forms of responsibility commonly considered in the literature: *counterfactual responsibility* and *responsibility for seeing to it*, respectively. In this paper, I study these two forms of responsibility in the multiagent systems that can be modelled as extensive form games.

The rest of this paper is divided into four major sections. In Section 2, I give a review of the two forms of responsibil-

[*] The full version of this paper includes an appendix, which can be found at arXiv:2312.07637.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

ity and related logic notions in the literature. Based on the discussion, I propose a new form of seeing-to-it responsibility for extensive form game settings in Section 3. Then, I formally define the model of extensive form games in Section 4 and the syntax and semantics of the two forms of responsibility in Section 5. In particular, I show the mutual undefinability between the two forms of responsibility, discuss the meaning of higher-order responsibility, and state the complexity of model checking. Finally, in Section 6, I formally study the *responsibility gap*. Although discussion of the responsibility gap is prevalent in the literature (Matthias 2004; Braham and VanHees 2011; Duijf 2018; Burton et al. 2020; Gunkel 2020; Langer et al. 2021; Goetze 2022), only a few studies (Braham and van Hees 2018; Hiller, Israel, and Heitzig 2022) give a formal definition of the concept. I then define the hierarchy of responsibility gaps, which, as far as I know, have never been discussed before. I show that a higher-order responsibility gap does not exist for sufficiently high orders.

2 Literature Review and Notion Discussion

Counterfactual responsibility captures the *principle of alternative possibilities* (Frankfurt 1969; Belnap and Perloff 1992; Widerker 2017): *an agent is responsible for a statement φ in an outcome if φ is true in the outcome and the agent had a strategy that could prevent it*. For example, consider outcome w_3 in Figure 1a, where the Board recommends (E) clemency and the Governor grants (G) it. In this case, Beach is set free. Note that both the Board and the Governor have a strategy (action D for the Board, action F for the Governor) to prevent this. As a result, each of them is counterfactually responsible for the fact that Beach, who was found by the court to be the murderer of Kim Nees, escapes punishment.

Next, consider outcome w_2 in which the Board recommends (E) clemency, but the Governor does not grant (F) it. Beach is left in prison. In this case, the Board is not counterfactually responsible for the fact that Beach is left in prison because the Board had no strategy to prevent this. At the same time, the Governor had such a strategy (action G). As a result, the Governor is counterfactually responsible for the fact that Beach is left in prison in outcome w_2.

Note that in order for Beach to be freed, both the Governor and the Board must support this. However, from the point of view of ascribing counterfactual responsibility, the order in which the decisions are made is important. If the Governor acts first, then, essentially, the roles of the Governor and the Board switch, see Figure 1b. In this new situation, the Governor is no longer counterfactually responsible for Beach being left in prison because he no longer has a strategy to prevent this. The dependency on the order of the decisions makes counterfactual responsibility in extensive form games different from the previously studied counterfactual responsibility in *strategic* game settings (Lorini and Schwarzentruber 2011; Naumov and Tao 2019, 2020), where all agents act concurrently and just once. The above definition of counterfactual responsibility for extensive form games is introduced in (Yazdanpanah et al. 2019). It also appears in (Baier, Funke, and Majumdar 2021).

The other commonly studied form of responsibility is defined through the notion of *seeing-to-it*. As a modality, seeing-to-it has been well studied in STIT logic (Chellas 1969; Belnap and Perloff 1990; Horty 2001; Horty and Pacuit 2017). Informally, an agent sees to it that φ if the agent guarantees that φ happens. When using the notion of seeing-to-it to define a form of responsibility, a *negative condition*[1] is usually required to exist to capture the intuition that no agent should be responsible for a trivial truth such as "$1 + 1 = 2$". The notion of *deliberative* seeing-to-it (Horty and Belnap 1995; Xu 1998; Balbiani, Herzig, and Troquard 2008; Olkhovikov and Wansing 2019) captures this idea by adding the requirement of a negative condition. Some follow-up work such as (Lorini, Longin, and Mayor 2014) and (Abarca and Broersen 2022) further incorporates the epistemic states of the agents into their discussion, but this is still within the STIT frame. Naumov and Tao (2021) studied deliberative seeing-to-it as one of the forms of responsibility in strategic game settings.

In extensive form game settings, there are two versions of the notion of seeing-to-it that may *potentially* capture a form of responsibility: *strategically* seeing-to-it in the presence of a negative condition and *achievement* seeing-to-it.

Strategically seeing-to-it (Broersen, Herzig, and Troquard 2006; Broersen 2009) is defined under the assumption that each agent commits upfront to a strategy (*i.e.* a plan of actions) for the duration of the game. Instead of guaranteeing φ to happen with one action, such a strategy guarantees φ to happen *in the final outcome* after acting according to the strategy in the whole game, no matter how the other agents may act in the process. For example, in the game depicted in Figure 1a, both the Board and the Governor have an upfront strategy to leave Beach in prison. For the Board, the strategy consists in denying the petition. For the Governor, the strategy consists in waiting for the Board to act and, if the Board recommends clemency, rejecting the petition. By incorporating the notion of a strategy, strategically seeing-to-it in the presence of a negative condition can be treated as a natural extension of deliberative seeing-to-it in multi-step decision schemes such as extensive form games.

However, this "natural extension" does not work for two reasons. On the one hand, by definition, the notion of strategically seeing-to-it has to be evaluated based on strategies rather than outcomes (Broersen and Herzig 2015). However, in some applications, such strategies may not be observable. Let us consider the case of outcome w_1 in Figure 1a. Here, the strategy of the Governor is not observable because he has no chance to make any choice. No one except for the Governor himself can tell how he would choose if the Board had not denied the clemency. Hence, even though he has a strategy to guarantee Beach being left in prison and the strategy is followed *in a trivial way* in outcome w_1, it is still not clear whether the Governor strategically sees to Beach being left in prison or not. On the other hand, although the strategy can be observed in some cases (such as pre-programmed au-

[1] In the general STIT models, a negative condition is a history where $\neg\varphi$ is true (Perloff 1991). In the extensive form game settings, a negative condition is an *outcome* where $\neg\varphi$ is true.

tonomous agents), the notion of strategically seeing-to-it accuses the agents of *thoughtcrime* purely based on their plans rather than actions. Note that, in law, *actus reus* ("guilty actions") is a commonly required element of a crime (Edwards 2021). For this reason, even if the Governor's strategy is to deny the clemency when the Board recommends it, which indeed strategically sees to Beach being left in prison according to the definition, the Governor should not be held *responsible* for seeing to this in outcome w_1, since he takes no action at all. Therefore, *strategically seeing-to-it in the presence of a negative condition* cannot always serve as a proper notion of responsibility in extensive form games.

Another notion of seeing-to-it that may capture a form of responsibility in extensive form game settings is *achievement seeing-to-it* (Belnap and Perloff 1992; Horty and Belnap 1995). In an extensive form game, the agents make choices one after another. Each choice of the agents may eliminate the possibility of some outcomes until the final outcome remains. If a statement is true in the final outcome, then during the game process, all the negative conditions, if exist, are eliminated. Achievement seeing-to-it captures the idea that, in such multi-step decision schemes, one specific *choice* of an agent guarantees some statement to be true in the final outcome by eliminating the "last possibility" for a negative condition to be achieved. For example, in outcome w_3 of Figure 1a, Beach is set free after the Board recommends (E) the clemency and the Governor grants (G) it. The choice of the Board (action E) eliminates one possibility of a negative condition (w_1) and the choice of the Governor (action G) eliminates the other possibility of a negative condition (w_2), which is also the last possibility. Hence, the Governor sees to it that Beach is set free in the achievement way in outcome w_3. Note that the notion of achievement seeing-to-it implies the existence of a negative condition by itself.

Achievement seeing-to-it can be treated as a form of responsibility in an intuitive sense. However, this notion cannot capture the idea of "guaranteeing" when regarding the extensive form game as a whole process. Let us still consider outcome w_3 in Figure 1a. When we treat the executive clemency procedure as a whole, the Governor does *not* guarantee that Beach will be set free, since the Board could have chosen to deny (D) the clemency before the Governor can make any decision. In fact, the Governor does not even have the *ability* to guarantee that Beach will be set free. Therefore, it is hard to say that the Governor is responsible for "seeing to it that" Beach is set free in outcome w_3, even though he sees to this in the achievement way.

The inconsistency between the notion of achievement seeing-to-it and the seeing-to-it form of responsibility is more significant when *obligation* is taken into consideration. For example, the obligation of doctors is to try their best to cure their patients. Consider a situation where a patient in danger of life is waiting for treatment. Suppose the treatment is sure to cure the patient. But the doctor leaves the patient unattended for six days and gives treatment on the seventh day. Then, the patient is cured. By giving the treatment, the doctor sees to it that the patient is cured in the achievement way. However, the doctor cannot be said to "be responsible (praiseworthy) for seeing to it that" the patient would be cured, because the patient might have died at any time during the first six days. For this reason, *achievement seeing-to-it* often cannot serve as a proper notion of the responsibility for seeing to it in extensive form games.

3 Responsibility for Seeing To It

In this section, I introduce a new notion of seeing-to-it responsibility that fits into extensive form game settings.

First, I modify the notion of strategically seeing-to-it into a backward version. I would say that an agent *backwards-strategically* sees to φ if the agent has an upfront ability to guarantee that φ would be true in the outcome and maintains the ability for the duration of the game. The ability to guarantee φ is captured by the *existence of a strategy* that guarantees φ. Note that, although the maintenance of the ability can be achieved by following such a strategy, the backward version of strategically seeing-to-it does *not* require the actually applied strategy to guarantee φ. Intuitively, instead of caring about the *plan* of the agent to guarantee φ, the backward version of strategically seeing-to-it focuses on the *ability* of guaranteeing it.

Unlike the original notion of strategically seeing-to-it, the backward version can be evaluated based on the outcomes (the paths of play) in extensive form game settings. For example, observe that in the game depicted in Figure 1a, the Board has the ability (the existence of a strategy) to guarantee that Beach would be left in prison at the b-labelled node and outcomes w_1 and w_2. The Governor has the same ability at the b-labelled node, the g-labelled node, and outcomes w_1 and w_2. On the path of play toward outcome w_1, both the Board and the Governor maintain this ability. Hence, in outcome w_1, both the Board and the Governor backwards-strategically see to Beach being left in prison. Note that w_1 is the outcome when the Governor applies the strategy "to grant (G) the clemency if the Board recommend (E) it" and the Board applies the strategy "to deny (D) the clemency". The Governor's strategy does *not* strategically see to Beach being left in prison in the original meaning. However, he still backwards-strategically sees to it. In outcome w_2, only the Governor backwards-strategically sees to Beach being left in prison because the Board loses the ability at the g-labelled node, where the Governor can grant (G) the clemency.

Second, I use the notion of backwards-strategically seeing-to-it, in combination with achievement seeing-to-it, to define the seeing-to-it form of responsibility in extensive form game settings. I would say that an agent is *responsible* for seeing to φ if she sees to it both backwards-strategically and in the achievement way. This combination captures both the ability and the action to "guarantee" in the notion of seeing-to-it. Informally, in the extensive form games, I say that an agent is responsible for seeing to φ if the agent *has an upfront ability to achieve φ, maintains it throughout the game, and eliminates the last possibility of a negative condition in the process*.

Consider the game depicted in Figure 1a. In outcome w_1, the Board sees to Beach being left in prison both backwards-strategically and in the achievement way. Therefore, the Board is responsible for seeing to Beach being left in prison in w_1. This argument is also true for the Governor in outcome w_2. However, in w_1, the Governor sees to Beach being left in prison backwards-strategically but not in the achievement way. Thus, the Governor is not responsible for seeing to this in w_1. In w_2, the Board sees to Beach being left in prison neither backwards-strategically nor in the achievement way. Hence, the Board is not responsible for seeing to this in w_2. Moreover, in outcome w_3, the Governor sees to Beach being set free in the achievement way but not backwards-strategically (he does not have such an ability at the b-labelled node). Thus, the governor is not responsible for seeing to Beach being set free in outcome w_3.

4 Extensive Form Games Terminology

In this section, I introduce extensive form games that are used later to give formal semantics of the modal language. Throughout the paper, I assume a fixed set of agents $\mathcal{A}$ and a fixed nonempty set of propositional variables.

Definition 1 *An extensive form game is a nonempty finite rooted tree in which each non-leaf node is labelled with an agent and each leaf node is labelled with a set of propositional variables.*

The leaf nodes of a game are called *outcomes* of the game. The set of all outcomes of a game G is denoted by $\Omega(G)$. An outcome is said to be labelled with a propositional variable if the outcome is labelled with a set containing this propositional variable. By $parent(n)$, I mean the parent node of any non-root node n. I write $n_1 \preceq n_2$ if node n_2 is on the simple path (including ends) between the root node and node n_1.

Definition 2 *For any set X of outcomes and any agent a, non-root node n is an X-achievement point by agent a, if*

1. *$parent(n)$ is labelled with agent a;*
2. *$w \notin X$ for some outcome w such that $w \preceq parent(n)$;*
3. *$w \in X$ for each outcome w such that $w \preceq n$.*

The notion of achievement point captures the idea that outcomes in set X are already "achieved" by agent a at node n: agent a choosing n at node $parent(n)$ eliminates the *last* possibility for an outcome *not* in X to be realised and thus guarantees that the game will end in X. For example, in the extensive form game depicted in Figure 1a, consider the set $\{w_1, w_2\}$ of outcomes where Beach is left in prison. Node w_1 is a $\{w_1, w_2\}$-achievement point by the Board, where action D of the Board at the b-labelled node eliminates the last possibility for Beach being set free (w_3) to come true. Similarly, node w_2 is a $\{w_1, w_2\}$-achievement point by the Governor. Note that an achievement point can also be a non-leaf node. For instance, the g-labelled node is a $\{w_2, w_3\}$-achievement point by the Board, since action E of the Board at the b-labelled node eliminates the last possibility for outcome w_1 to be realised. The next lemma shows a property of achievement point, whose significance is due to the uniqueness of the chance to eliminate the last possibility.

Lemma 1 *For any extensive form game G, any set of outcomes $X \subsetneq \Omega(G)$, and any outcome $w \in X$, there is a unique agent a and a unique X-achievement point n by agent a such that $w \preceq n$.*

Next, I define the notation $win_a(X)$. For any set X of outcomes and any agent a, by $win_a(X)$ I mean the set of all nodes (including outcomes) from which agent a has the *ability* to end the game in set X. Formally, the set $win_a(X)$ is defined below using backward induction.

Definition 3 *For any set X of outcomes, the set $win_a(X)$ is the minimal set of nodes such that*

1. *$X \subseteq win_a(X)$;*
2. *for any non-leaf node n labelled with agent a, if **at least one** child of node n belongs to the set $win_a(X)$, then node $n \in win_a(X)$;*
3. *for any non-leaf node n **not** labelled with agent a, if **all** children of node n belong to the set $win_a(X)$, then node $n \in win_a(X)$.*

Informally, for a non-leaf node $n \in win_a(X)$ labelled with agent a, the ability of agent a to end the game in X is captured by the strategy that always chooses a child node of n from the set $win_a(X)$. Specifically, if the root of the tree is in the set $win_a(X)$, then agent a has an upfront ability to end the game in X.

5 Syntax and Semantics

The language Φ of the logical system is defined by the following grammar:

$$\varphi := p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \mathsf{C}_a\varphi \mid \mathsf{S}_a\varphi,$$

where p is a propositional variable and $a \in \mathcal{A}$ is an agent. The formula $\mathsf{C}_a\varphi$ is read as "agent a is counterfactually responsible for φ" and $\mathsf{S}_a\varphi$ is read as "agent a is responsible for seeing to φ". Boolean constants true $\top$ and false $\bot$ are defined in the standard way.

The next is the core definition of this paper. Informally, for each formula $\varphi \in \Phi$, the truth set $[\![\varphi]\!]$ is the set of all outcomes where φ is true.

Definition 4 *For any extensive form game G and any formula $\varphi \in \Phi$, the truth set $[\![\varphi]\!]$ is defined recursively:*

1. *$[\![p]\!]$ is the set of all outcomes labelled with propositional variable p;*
2. *$[\![\neg\varphi]\!] = \Omega(G) \setminus [\![\varphi]\!]$;*
3. *$[\![\varphi \wedge \psi]\!] = [\![\varphi]\!] \cap [\![\psi]\!]$;*
4. *$[\![\mathsf{C}_a\varphi]\!]$ is the set of all outcomes $w \in [\![\varphi]\!]$ such that there exists a node $n \in win_a([\![\neg\varphi]\!])$ where $w \preceq n$;*
5. *$[\![\mathsf{S}_a\varphi]\!]$ is the set of all outcomes $w \in \Omega(G)$ such that*
   (a) *$\{n \mid w \preceq n\} \subseteq win_a([\![\varphi]\!])$;*
   (b) *there exists a $[\![\varphi]\!]$-achievement point n by agent a such that $w \preceq n$.*

Item 4 above defines the notion of counterfactual responsibility following (Yazdanpanah et al. 2019) and (Baier, Funke, and Majumdar 2021). An agent a is counterfactually responsible for a statement φ in outcome w if two conditions are satisfied: (i) φ is true in w and (ii) on the path of play, agent a has a strategy to prevent φ. The first condition is captured by the assumption $w \in [\![\varphi]\!]$. The second condition is captured by the existence of a node n on the path of play ($w \preceq n$) to outcome w such that $n \in win_a([\![\neg\varphi]\!])$.

Item 5 above defines the seeing-to-it form of responsibility as the combination of backwards-strategically seeing-to-it and achievement seeing-to-it. An agent backwards-strategically sees to φ in outcome w if the agent has an upfront ability to achieve φ and maintains the ability throughout the game. This is captured by the fact that all the nodes on the path of play leading to outcome w belong to the set $win_a(\llbracket\varphi\rrbracket)$, as part 5(a) of Definition 4 shows. An agent sees to φ in the achievement way in outcome w if the agent eliminates the last possibility for $\neg\varphi$. This means, on the path of play toward outcome w, there exists a $\llbracket\varphi\rrbracket$-achievement point by agent a. This is captured in part 5(b) of Definition 4.

5.1 Mutual Undefinability Between C and S

Modalities C and S capture two forms of responsibility in extensive form games. One natural question is: are they both needed? The answer would be no if one of them can be defined via the other. For example, in propositional logic, implication $\to$ and disjunction $\vee$ can be defined via negation $\neg$ and conjunction $\wedge$.[2] Hence, using only negation $\neg$ and conjunction $\wedge$ is enough for a propositional logic system. Note that Naumov and Tao (2021) proved that modality S is not definable via modality C but modality C is definable via modality S by $\mathsf{C}_a\varphi \equiv \varphi \wedge \mathsf{S}_a\neg\mathsf{S}_a\neg\varphi$ in *strategic game settings*. Before presenting the results about extensive form games, let me start with an auxiliary definition:

Definition 5 *Formulae $\varphi, \psi \in \Phi$ are semantically equivalent if $\llbracket\varphi\rrbracket = \llbracket\psi\rrbracket$ for each extensive form game.*

In language Φ, modality C is *definable* via modality S if, for each formula $\varphi \in \Phi$, there is a semantically equivalent formula $\psi \in \Phi$ that does *not* use modality C. The definability of S via C could be specified similarly. The two theorems below show that C and S are *not* definable via each other in extensive form game settings. These results show that, in order to discuss both forms of responsibility in extensive form game settings, both modalities are needed.

Theorem 1 (undefinability of C via S) *The formula $\mathsf{C}_a p$ is not semantically equivalent to any formula in language Φ that does not contain modality C.*

Theorem 2 (undefinability of S via C) *The formula $\mathsf{S}_a p$ is not semantically equivalent to any formula in language Φ that does not contain modality S.*

Using the "truth set algebra" technique (Knight et al. 2022), I formally proof the above two theorems in Appendices A.1 and A.2 of the full version of this paper (Shi 2023).

5.2 Higher-Order Responsibility

As can be seen in the grammar of language Φ, modalities C and S can be nested in a formula. By higher-order responsibility, I mean more complicated forms of responsibility expressible by the nesting of modalities C and S. For example, in outcome w_2 of the game depicted in Figure 1a, the Governor is counterfactually responsible for Beach being left in prison. However, the Board could have *prevented such responsibility* by denying (D) the petition. Thus, in outcome w_2, the Board is counterfactually responsible for the Governor's responsibility for Beach being left in prison: $w_2 \in \llbracket\mathsf{C}_b\mathsf{C}_g\text{“Beach is left in prison”}\rrbracket$. Similarly, it is true that $w_2 \in \llbracket\mathsf{C}_b\mathsf{S}_g\text{“Beach is left in prison”}\rrbracket$.

Discussion of higher-order responsibility makes sense, especially in a situation where some of the agents who do affect the outcome are not the *proper subjects* to ascribe the responsibility. For example, young kids are usually not considered the proper subjects of criminal responsibility. Therefore, when they commit crimes and assume direct responsibility for the outcomes, the secondary responsibility of their guardians needs to be considered (Hollingsworth 2007). The same is true for autonomous agents and their designers.

There are some interesting properties of higher-order responsibility. For instance, *formulae $\mathsf{C}_a\mathsf{C}_a\varphi$ and $\mathsf{C}_a\varphi$ are semantically equivalent* (see Property 1 in Appendix B.1 of the full version). This means, if an agent is counterfactually responsible for a statement φ, then she is also counterfactually responsible for assuming this counterfactual responsibility. Also, *both formulae $\mathsf{S}_b\mathsf{S}_a\varphi$ and $\mathsf{S}_b\mathsf{C}_a\varphi$ are semantically equivalent to $\perp$* (see Property 2 and Property 3 in Appendices B.2 and B.3 of the full version). This means an agent can never be responsible for seeing to the responsibility of another agent.

5.3 Complexity of Model Checking

In the setting of this paper, the computation of the set $\llbracket\varphi\rrbracket$ is the core of any model checking problem related to formula φ. Hence, I analyse the complexity of computing the truth set $\llbracket\varphi\rrbracket$ for an arbitrary formula $\varphi \in \Phi$. Assume that deciding whether an outcome is labelled with a propositional variable takes constant time. Then, the next theorem follows from Definition 2, Definition 3, and Definition 4. See Appendix C of the full version for detailed analysis.

Theorem 3 (time complexity) *For any formula $\varphi \in \Phi$ and any extensive form game G, the computation of the set $\llbracket\varphi\rrbracket$ takes $O(|\varphi| \cdot |G|)$, where $|\varphi|$ is the size of formula φ and $|G|$ is the number of nodes in game G.*

6 Responsibility Gap

One of the important questions discussed in the ethics literature is the responsibility gap. That is, if something happens, is there always an agent that can be held responsible for it? I now discuss if the two forms of responsibility considered in the paper are enough to avoid responsibility gaps in extensive form games. Note that, as I discussed in Section 2, nobody should be responsible for a vacuous truth. Hence, in this section, I only consider the responsibility gaps for statements that are *not* trivially true.

Let us go back to the example depicted in Figure 1a. Recall that if Beach is left in prison, then in outcome w_1, the Board is responsible for seeing to it; in outcome w_2, the Governor is responsible for seeing to it and also counterfactually responsible for it. If Beach is set free in outcome w_3, then the Governor is counterfactually responsible for it. Thus, for the statements "Beach is left in prison" and "Beach is set free", there is no responsibility gap in this game.

[2] $\varphi \to \psi \equiv \neg(\varphi \wedge \neg\psi)$ and $\varphi \vee \psi \equiv \neg(\neg\varphi \wedge \neg\psi)$ for each formulae φ, ψ in propositional logic.

6.1 In Extensive Form Games With Two Agents

In Theorem 4 below, I show that the two forms of responsibility discussed in this paper leave no gap in any extensive form games with only two agents. Let us start, however, by formally defining the ***gap formulae*** $\mathsf{G}^{\mathsf{c}}(\varphi)$ and $\mathsf{G}^{\mathsf{s}}(\varphi)$ for any formula $\varphi \in \Phi$. Informally, the formula $\mathsf{G}^{\mathsf{c}}(\varphi)$ means that φ is true and nobody is counterfactually responsible for it. The formula $\mathsf{G}^{\mathsf{s}}(\varphi)$ says the same for the seeing-to-it form of responsibility.

$$\mathsf{G}^{\mathsf{c}}(\varphi) := \varphi \wedge \bigwedge_{a\in\mathcal{A}} \neg\mathsf{C}_a\varphi, \tag{1}$$

$$\mathsf{G}^{\mathsf{s}}(\varphi) := \varphi \wedge \bigwedge_{a\in\mathcal{A}} \neg\mathsf{S}_a\varphi. \tag{2}$$

The combined responsibility gap formula $\mathsf{G}^{\mathsf{c,s}}(\varphi)$ is defined as the conjunction $\mathsf{G}^{\mathsf{c}}(\varphi) \wedge \mathsf{G}^{\mathsf{s}}(\varphi)$. The proof of Theorem 4 uses the following well-known lemma. To keep this paper self-contained, I prove the lemma in Appendix D.1 of the full version.

Lemma 2 *For any formula $\varphi \in \Phi$ and any node n in a two-agent extensive form game between agents a and b, if $n \notin win_a(\llbracket\varphi\rrbracket)$, then $n \in win_b(\llbracket\neg\varphi\rrbracket)$.*

Theorem 4 *For any formula $\varphi \in \Phi$ and any two-agent extensive form game G, if $\llbracket\varphi\rrbracket \neq \Omega(G)$, then $\llbracket\mathsf{G}^{\mathsf{c,s}}(\varphi)\rrbracket = \varnothing$.*

PROOF. I prove this theorem by showing that, for any outcome $w \in \Omega(G)$, if $w \in \llbracket\mathsf{G}^{\mathsf{s}}(\varphi)\rrbracket$, then $w \notin \llbracket\mathsf{G}^{\mathsf{c}}(\varphi)\rrbracket$. By statement (1) and items 2 and 3 of Definition 4, it suffices to show the existence of an agent a such that $w \in \llbracket\mathsf{C}_a\varphi\rrbracket$.

By statement (2) and items 2 and 3 of Definition 4, the assumption $w \in \llbracket\mathsf{G}^{\mathsf{s}}(\varphi)\rrbracket$ implies that

$$w \in \llbracket\varphi\rrbracket \tag{3}$$

and

$$w \notin \llbracket\mathsf{S}_b\varphi\rrbracket \tag{4}$$

for each agent $b \in \mathcal{A}$. At the same time, by the assumption $\llbracket\varphi\rrbracket \neq \Omega(G)$, statement (3), and Lemma 1, there exists an agent b and a $\llbracket\varphi\rrbracket$-achievement point n by agent b such that $w \preceq n$. Hence, by item 5 of Definition 4 and statement (4), there is a node m such that $w \preceq m$ and $m \notin win_b(\llbracket\varphi\rrbracket)$. Since G is a two-agent game, let a be the agent in the game distinct from agent b. Then, $m \in win_a(\llbracket\neg\varphi\rrbracket)$ by Lemma 2. Hence, $w \in \llbracket\mathsf{C}_a\varphi\rrbracket$ by item 4 of Definition 4 because of statement (3) and that $w \preceq m$. □

6.2 In Extensive Form Games With More Agents

To see if there is a responsibility gap in extensive form games with *more than* two agents, let us go back to the story in the introduction, which is not as simple as I tried to make it. In over 30 years that separate Kim Nees's murder and Beach's attorney filing an executive clemency petition, the case became highly controversial in Montana due to the lack of direct evidence and doubts about the integrity of the interrogators. By the time the petition was filed, the Board had already made clear its intention to deny the petition, while the Governor expressed his support for the clemency (Bullock 2015).

Then, something very unusual happened. On 4 December 2014, a bill was introduced in the Montana House of Representatives that would allow the Governor to grant executive clemency no matter what the decision of the Board is. This bill aimed to strip the Board from the power that it had from the day the State of Montana was founded in 1889. Although the bill would affect the Governor's power to grant clemency in other cases as well, the primary goal of the legislation was to give the Governor a chance to free Beach (Montana Innocence Project 2023).

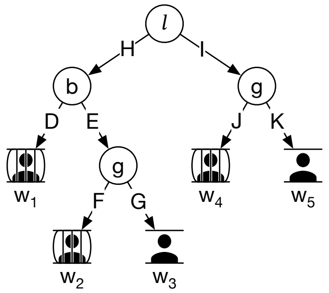


Figure 2: Barry Beach's case of clemency

Figure 2 depicts the extensive form game that captures the situation after the bill was introduced. If the Montana State Legislature rejects (H) the bill, then the game continues as in Figure 1a. If the Legislature approves (I) the bill, then the Governor unilaterally decides whether to grant the clemency. In this new three-agent game, the Governor is responsible for Beach being left in prison in outcomes w_2 and w_4 both counterfactually and for seeing to it. The Governor is also counterfactually responsible for Beach being freed in outcomes w_3 and w_5. However, in outcome w_1, nobody is responsible for the fact that Beach is left in prison either counterfactually or for seeing to it. In particular, the Board is not responsible for seeing to this because it no longer has an upfront ability to guarantee that Beach is left in prison in the outcome. Therefore, by statements (1) and (2),

$$\llbracket\mathsf{G}^{\mathsf{c,s}}(\text{“Beach is left in prison”})\rrbracket = \{w_1\}. \tag{5}$$

This example shows that the responsibility gap may exist in extensive form games with more than two agents. In other words, the two forms of responsibility discussed here are not enough to have a responsible agent in every situation.

6.3 Hierarchy of Responsibility Gaps

A further question about the responsibility gap is if there is an agent responsible for the gap. The responsibility for the gap, or *the responsibility for the lack of a responsible agent*, is a natural concept that applies to many real-world situations. For instance, the managers who assign tasks and the governing bodies that set the rules are often responsible for the lack of a responsible person. In the example in Figure 2, it is the Legislature that is counterfactually responsible for the gap in outcome w_1. Indeed, the Legislature could prevent the formula $\mathsf{G}^{\mathsf{c,s}}(\text{“Beach is left in prison”})$ from being true by approving (I) the bill:

$$w_1 \in \llbracket\mathsf{C}_l\mathsf{G}^{\mathsf{c,s}}(\text{“Beach is left in prison”})\rrbracket.$$

In addition, in this example, the Board is also counterfactually responsible for the gap in outcome w_1.

I also consider the lack of responsibility for the gap. By *second-order gap* for a formula φ I mean the presence of outcomes in which $\mathsf{G}^{\mathsf{c,s}}(\varphi)$ is true and nobody is responsible for it. In a real-world situation, the first-order responsibility gap often shows that the managers do not assign tasks in an accountable way, while the second-order responsibility gap is often caused by a failure of the leadership to properly define the roles of the managers so that the managers had no way to assign tasks in an accountable way.

In general, for an arbitrary formula $\varphi \in \Phi$ and any integer $i \geq 0$, let the i^{th}-order gap statement $\mathsf{G}^{\mathsf{c,s}}_i(\varphi)$ be defined recursively as:

$$\mathsf{G}^{\mathsf{c,s}}_i(\varphi) := \begin{cases} \mathsf{G}^{\mathsf{c,s}}_{i-1}(\varphi) \wedge \bigwedge_{a\in\mathcal{A}} \neg\mathsf{C}_a\mathsf{G}^{\mathsf{c,s}}_{i-1}(\varphi) & \\ \qquad\qquad \wedge \bigwedge_{a\in\mathcal{A}} \neg\mathsf{S}_a\mathsf{G}^{\mathsf{c,s}}_{i-1}(\varphi), & i \geq 1; \\ \varphi, & i = 0. \end{cases} \tag{6}$$

In addition, let the i^{th}-order counterfactual gap statement $\mathsf{G}^{\mathsf{c}}_i(\varphi)$ be defined recursively as:

$$\mathsf{G}^{\mathsf{c}}_i(\varphi) := \begin{cases} \mathsf{G}^{\mathsf{c}}_{i-1}(\varphi) \wedge \bigwedge_{a\in\mathcal{A}} \neg\mathsf{C}_a\mathsf{G}^{\mathsf{c}}_{i-1}(\varphi), & i \geq 1; \\ \varphi, & i = 0. \end{cases} \tag{7}$$

One can similarly define the i^{th}-order seeing-to-it gap statement $\mathsf{G}^{\mathsf{s}}_i(\varphi)$. It is easy to see from statements (1) and (2) that the first order gap statements $\mathsf{G}^{\mathsf{c,s}}_1(\varphi)$, $\mathsf{G}^{\mathsf{c}}_1(\varphi)$, and $\mathsf{G}^{\mathsf{s}}_1(\varphi)$ are equivalent to the previously discussed gap statement $\mathsf{G}^{\mathsf{c,s}}(\varphi)$, $\mathsf{G}^{\mathsf{c}}(\varphi)$, and $\mathsf{G}^{\mathsf{s}}(\varphi)$, respectively.

As shown in Theorem 4, in two-agent extensive form games, the truth set $\llbracket \mathsf{G}^{\mathsf{c,s}}(\varphi) \rrbracket$ is empty for each formula $\varphi \in \Phi$ such that $\llbracket\varphi\rrbracket \neq \Omega(G)$. Informally, this means that there is no responsibility gap in two-agent extensive form games. At the same time, the example depicted in Figure 2 shows that such a gap might exist in games with more than two agents. This observation is correct. In Appendix D.2 of the full version, for each integer $i \geq 2$, I construct an extensive form game in which the truth set $\llbracket \mathsf{G}^{\mathsf{c,s}}_i(\varphi) \rrbracket$ is *not* empty.

Despite this, in Theorem 5 and Corollary 1 below, I show that, for any extensive form game, the sets $\llbracket \mathsf{G}^{\mathsf{c}}_i(\varphi) \rrbracket$ and $\llbracket \mathsf{G}^{\mathsf{c,s}}_i(\varphi) \rrbracket$ are empty for large enough integer i. Informally, the higher-order responsibility gap does not exist in any extensive form game if sufficiently high order is considered.

Let me first show two lemmas that are used later to prove Theorem 5. These lemmas show that the set $\llbracket \mathsf{G}^{\mathsf{c}}_i(\varphi) \rrbracket$ monotonously shrinks to empty as the order i increases.

Lemma 3 $\llbracket \mathsf{G}^{\mathsf{c}}_{i+1}(\varphi) \rrbracket \subseteq \llbracket \mathsf{G}^{\mathsf{c}}_i(\varphi) \rrbracket$ *for any formula* $\varphi \in \Phi$ *and any integer* $i \geq 0$.

PROOF. The statement of the lemma follows from statement (7) and item 3 of Definition 4. □

Lemma 4 *For any formula* $\varphi \in \Phi$, *any integer* $i \geq 0$, *and any extensive form game* G, *if* $\varnothing \subsetneq \llbracket \mathsf{G}^{\mathsf{c}}_i(\varphi) \rrbracket \subsetneq \Omega(G)$, *then* $\llbracket \mathsf{G}^{\mathsf{c}}_{i+1}(\varphi) \rrbracket \subsetneq \llbracket \mathsf{G}^{\mathsf{c}}_i(\varphi) \rrbracket$.

PROOF. The assumption $\varnothing \subsetneq \llbracket \mathsf{G}^{\mathsf{c}}_i(\varphi) \rrbracket \subsetneq \Omega(G)$, by item 2 of Definition 4, implies that $\varnothing \subsetneq \llbracket \neg\mathsf{G}^{\mathsf{c}}_i(\varphi) \rrbracket \subsetneq \Omega(G)$. Then, on the one hand, there is an outcome $w \in \llbracket \neg\mathsf{G}^{\mathsf{c}}_i(\varphi) \rrbracket$. On the other hand, by Lemma 1, there is an $\llbracket \neg\mathsf{G}^{\mathsf{c}}_i(\varphi) \rrbracket$-achievement point n by an agent a such that $w \preceq n$. Thus, by Definition 2,

1. $parent(n)$ is labelled with agent a;
2. there exists an outcome w' such that $w' \preceq parent(n)$ and $w' \notin \llbracket \neg\mathsf{G}^{\mathsf{c}}_i(\varphi) \rrbracket$;
3. $w'' \in \llbracket \neg\mathsf{G}^{\mathsf{c}}_i(\varphi) \rrbracket$ for each outcome w'' such that $w'' \preceq n$.

Item 3 above implies that $n \in win_a(\llbracket \neg\mathsf{G}^{\mathsf{c}}_i(\varphi) \rrbracket)$ by Definition 3. Hence, by item 2 of Definition 3 and item 1 above,

$$parent(n) \in win_a(\llbracket \neg\mathsf{G}^{\mathsf{c}}_i(\varphi) \rrbracket). \tag{8}$$

By the part $w' \notin \llbracket \neg\mathsf{G}^{\mathsf{c}}_i(\varphi) \rrbracket$ of item 2 above and item 2 of Definition 4,

$$w' \in \llbracket \mathsf{G}^{\mathsf{c}}_i(\varphi) \rrbracket. \tag{9}$$

Thus, $w' \in \llbracket \mathsf{C}_a\mathsf{G}^{\mathsf{c}}_i(\varphi) \rrbracket$ by the part $w' \preceq parent(n)$ of item 2 above, statement (8), and item 4 of Definition 4. Then, $w' \notin \llbracket \neg\mathsf{C}_a\mathsf{G}^{\mathsf{c}}_i(\varphi) \rrbracket$ by item 2 of Definition 4. Hence, $w' \notin \llbracket \mathsf{G}^{\mathsf{c}}_{i+1}(\varphi) \rrbracket$ by statement (7) and item 3 of Definition 4. Then, $\llbracket \mathsf{G}^{\mathsf{c}}_{i+1}(\varphi) \rrbracket \neq \llbracket \mathsf{G}^{\mathsf{c}}_i(\varphi) \rrbracket$ by statement (9). Therefore, $\llbracket \mathsf{G}^{\mathsf{c}}_{i+1}(\varphi) \rrbracket \subsetneq \llbracket \mathsf{G}^{\mathsf{c}}_i(\varphi) \rrbracket$ by Lemma 3. □

Theorem 5 $\llbracket \mathsf{G}^{\mathsf{c}}_i(\varphi) \rrbracket = \varnothing$ *for each integer* $i \geq |\Omega(G)| - 1$ *and each formula* $\varphi \in \Phi$ *such that* $\llbracket\varphi\rrbracket \subsetneq \Omega(G)$.

PROOF. By the assumption $\llbracket\varphi\rrbracket \subsetneq \Omega(G)$ of this theorem, statement (7), and Lemma 3,

$$\Omega(G) \supsetneq \llbracket\varphi\rrbracket = \llbracket \mathsf{G}^{\mathsf{c}}_0(\varphi) \rrbracket \supseteq \llbracket \mathsf{G}^{\mathsf{c}}_1(\varphi) \rrbracket \supseteq \llbracket \mathsf{G}^{\mathsf{c}}_2(\varphi) \rrbracket \supseteq \dots$$

Note that $|\llbracket\varphi\rrbracket| \leq |\Omega(G)| - 1$ by the assumption $\llbracket\varphi\rrbracket \subsetneq \Omega(G)$ of this theorem. Therefore, $\llbracket \mathsf{G}^{\mathsf{c}}_i(\varphi) \rrbracket = \varnothing$ for each integer $i \geq |\Omega(G)| - 1$ by Lemma 4. □

The next corollary follows from the above theorem and the observation that $\llbracket \mathsf{G}^{\mathsf{c,s}}_i(\varphi) \rrbracket \subseteq \llbracket \mathsf{G}^{\mathsf{c}}_i(\varphi) \rrbracket$. I give the formal proof in Appendix D.3 of the full version.

Corollary 1 $\llbracket \mathsf{G}^{\mathsf{c,s}}_i(\varphi) \rrbracket = \varnothing$ *for each integer* $i \geq |\Omega(G)| - 1$ *and each formula* $\varphi \in \Phi$ *such that* $\llbracket\varphi\rrbracket \subsetneq \Omega(G)$.

7 Conclusion

The existing definitions of seeing-to-it modalities have clear shortcomings when viewed as possible forms of responsibility. In this paper, I combined them into a single definition of seeing-to-it responsibility that addresses the shortcomings. By proving the undefinability results, I have shown that the proposed notion is semantically independent of the counterfactual responsibility already discussed in the literature. The other important contribution of this work is the hierarchy of responsibility gaps. I believe that taking into account higher-order responsibilities is an important step towards designing better mechanisms in terms of responsibility attribution. In the future, I would like to study how the gap results could be extended to the setting of games with imperfect information, where even Lemma 2 does not hold.

One more thing, if you are curious about the ending of Beach's story, in January 2015, the Montana House of Representatives approved the bill that changes the clemency procedure. By doing so, they, perhaps unintentionally, prevented the potential responsibility gap existing in outcome w_1 of Figure 2. In November of the same year, the Governor granted clemency to Beach (Bullock 2015).

Acknowledgements

I would like to express my sincere gratitude to Dr. Pavel Naumov, my PhD supervisor, for his selfless help in completing this paper. The research is funded by the China Scholarship Council (CSC No.202206070014).

References

Abarca, A. I. R.; and Broersen, J. M. 2022. A STIT logic of responsibility. In *Proceeding of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS-22)*, 1717–1719.

Associated Press. 2015. Montana governor frees man convicted in 1979 beating death of classmate. *The Guardian*, November 20. https://www.theguardian.com/us-news/2015/nov/20/montana-governor-grants-clemency-barry-beach.

Baier, C.; Funke, F.; and Majumdar, R. 2021. A game-theoretic account of responsibility allocation. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI-21)*, 1773–1779.

Balbiani, P.; Herzig, A.; and Troquard, N. 2008. Alternative axiomatics and complexity of deliberative STIT theories. *Journal of Philosophical Logic*, 37(4): 387–406.

Belnap, N.; and Perloff, M. 1990. Seeing to it that: a canonical form for agentives. In *Knowledge Representation and Defeasible Reasoning*, 167–190. Springer.

Belnap, N.; and Perloff, M. 1992. The way of the agent. *Studia Logica*, 51: 463–484.

Braham, M.; and van Hees, M. 2018. Voids or fragmentation: moral responsibility for collective outcomes. *The Economic Journal*, 128(612): F95–F113.

Braham, M.; and VanHees, M. 2011. Responsibility voids. *The Philosophical Quarterly*, 61(242): 6–15.

Broersen, J. 2009. A STIT-logic for extensive form group strategies. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, volume 3, 484–487. IEEE.

Broersen, J.; and Herzig, A. 2015. Using STIT theory to talk about strategies. *Models of Strategic Reasoning: Logics, Games, and Communities*, 137–173.

Broersen, J.; Herzig, A.; and Troquard, N. 2006. A STIT-extension of ATL. In *European Workshop on Logics in Artificial Intelligence*, 69–81. Springer.

Bullock, S. 2015. Executive Order Granting Clemency to Barry Allan Beach. https://formergovernors.mt.gov/bullock/docs/2015EOs/EO_19_2015_Beach.pdf. Accessed: 2023-05-14.

Burton, S.; Habli, I.; Lawton, T.; McDermid, J.; Morgan, P.; and Porter, Z. 2020. Mind the gaps: assuring the safety of autonomous systems from an engineering, ethical, and legal perspective. *Artificial Intelligence*, 279: 103201.

Chellas, B. F. 1969. *The logical form of imperatives*. Stanford University.

Constitution Convention. 1889. Constitution of the State of Montana. https://courts.mt.gov/external/library/docs/1889cons.pdf. Accessed: 2023-05-14.

Duijf, H. 2018. Responsibility voids and cooperation. *Philosophy of the Social Sciences*, 48(4): 434–460.

Edwards, J. 2021. Theories of criminal law. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

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

Goetze, T. S. 2022. Mind the gap: autonomous systems, the responsibility gap, and moral entanglement. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 390–400.

Gunkel, D. J. 2020. Mind the gap: responsible robotics and the problem of responsibility. *Ethics and Information Technology*, 22(4): 307–320.

Hiller, S.; Israel, J.; and Heitzig, J. 2022. An axiomatic approach to formalized responsibility ascription. In *Proceedings of the 24th International Conference on Principles and Practice of Multi-Agent Systems (PRIMA-22)*, 435–457. Springer.

Hollingsworth, K. 2007. Responsibility and rights: children and their parents in the youth justice system. *International Journal of Law, Policy and the Family*, 21(2): 190–219.

Horty, J.; and Pacuit, E. 2017. Action types in STIT semantics. *The Review of Symbolic Logic*, 10(4): 617–637.

Horty, J. F. 2001. *Agency and deontic logic*. Oxford University Press.

Horty, J. F.; and Belnap, N. 1995. The deliberative STIT: a study of action, omission, ability, and obligation. *Journal of Philosophical Logic*, 24(6): 583–644.

Knight, S.; Naumov, P.; Shi, Q.; and Suntharraj, V. 2022. Truth set algebra: a new way to prove undefinability. *arXiv:2208.04422*.

Langer, M.; Oster, D.; Speith, T.; Hermanns, H.; Kästner, L.; Schmidt, E.; Sesing, A.; and Baum, K. 2021. What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, 296: 103473.

Lorini, E.; Longin, D.; and Mayor, E. 2014. A logical analysis of responsibility attribution: emotions, individuals and collectives. *Journal of Logic and Computation*, 24(6): 1313–1339.

Lorini, E.; and Schwarzentruber, F. 2011. A logic for reasoning about counterfactual emotions. *Artificial Intelligence*, 175(3): 814–847.

Matthias, A. 2004. The responsibility gap: ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6: 175–183.

Montana Board of Pardons and Parole. 2023. History. https://bopp.mt.gov/History. Accessed: 2023-05-14.

Montana Innocence Project. 2023. Never, ever, ever give up: Barry Beach's resilient fight for freedom. https://mtinnocenceproject.org/barry-beach/. Accessed: 2023-05-26.

Naumov, P.; and Tao, J. 2019. Blameworthiness in strategic games. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI-19)*, 3011–3018.

Naumov, P.; and Tao, J. 2020. An epistemic logic of blameworthiness. *Artificial Intelligence*, 283: 103269.

Naumov, P.; and Tao, J. 2021. Two forms of responsibility in strategic games. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI-21)*.

Olkhovikov, G. K.; and Wansing, H. 2019. Inference as doxastic agency. Part I: The basics of justification STIT logic. *Studia Logica*, 107(1): 167–194.

Perloff, M. 1991. STIT and the language of agency. *Synthese*, 86: 379–408.

Shi, Q. 2023. Responsibility in extensive form games. *arXiv:2312.07637*.

US Supreme Court. 1993. Herrera v. Collins, 506 U.S. 390. https://supreme.justia.com/cases/federal/us/506/390. Accessed: 2023-12-23.

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.

Yazdanpanah, V.; Dastani, M.; Alechina, N.; Logan, B.; and Jamroga, W. 2019. Strategic responsibility under imperfect information. In *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS-19)*, 592–600.