Journal of Philosophical Logic
https://doi.org/10.1007/s10992-025-09794-7

Responsibility in Multi-Step Decision Schemes

Qi Shi[1] · **Pavel Naumov**[1]

Received: 24 May 2024 / Accepted: 24 March 2025

Abstract
Two different forms of responsibility, counterfactual and seeing-to-it, have been extensively discussed in philosophy 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 counterfactual responsibility and the newly proposed notion of responsibility are not definable via each other. It also studies the higher-order responsibility and 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.

Keywords Responsibility · STIT · Undefinability · Responsibility gap

1 Introduction

In this paper [1], we study responsibility in the context of collective decision-making. The notion of responsibility, in linguistic intuition, can be divided into two different classes: the ***forward***-perspective responsibility and the ***backward***-perspective responsibility [2]. The forward-perspective responsibility focuses on seeing to it that a certain state of affairs is obtained [3]. More specifically, it takes into account the eventualities as

[1] A preliminary version of this work appeared in AAAI-24 proceedings [1].

✉ Qi Shi
qi.shi@soton.ac.uk

Pavel Naumov
p.naumov@soton.ac.uk

[1] School of Electronics & Computer Science, University of Southampton, University Road, Southampton SO17 1BJ, UK

Published online: 21 April 2025

potential situations that may be materialised in the future and analyses how the agents can or ought to affect such state of affairs [4]. For example, the *role-responsibility* (*e.g.* a sea captain is responsible for the safety of his ship) [5] that talks about *obligations* is a type of forward-perspective responsibility. In contrast, the backward-perspective responsibility looks at the affairs that have already or hypothetically happened and is closely related the philosophical notions like *accountability*, *blameworthiness*, and *liability* [2]. The discussion of the backward-perspective responsibility is deeply rooted in the analysis of the causal chains, which is studied in the theory of *actual causality* [6], and the agency of the parties involved in the affairs [7]. Dastani and Yazdanpanah [7] call the backward-perspective responsibility by actual responsibility and, based on the difference in *methodology*, divide it into two types: ***event***-oriented responsibility [6, 8, 9] that uses causal models and treats the agents and their actions as the general events in the models and ***agent***-oriented responsibility [10–12] that uses strategic settings (*e.g.* games) and considers the strategic abilities and epistemic states of the agents in those settings. However, in terms of the *concept*, there is no borderline between event-oriented responsibility and agent-oriented responsibility. Moreover, the agent-oriented definitions of responsibility are usually the refinement of the event-oriented ones that take the agency of the agents into consideration.

The focus of this work is on the agent-oriented responsibility in multi-step decision schemes where the agents make decisions sequentially and their joint decision determines the final outcome. We model such schemes as ***extensive form games*** (see Definition 1) and use this term henceforth. Notice that, unlike some other researchers who are concerned with who or what can be held responsible [13–19], we treat all the subjects that have agency and can affect the outcome in a system as ***agents***, such as humans, animals, and artificial intelligence. Holding the opinion that whether an agent is a *proper subject* to ascribe responsibility[2] does not affect the fact that the agent is responsible (*e.g.* from the perspective of the causal chain), we discuss the responsibility for the general notion of "agent". It is also worth mentioning that, by using the extensive form game model, we assume the decision scheme is fixed and known to every agent in it. By this means, the control condition and epistemic condition to hold an agent responsible [20] are both met.

Following the 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, as well as the fact that the agent can act to prevent or to achieve a certain result, in this paper, we consider two forms of backward-perspective agent-oriented responsibility that have been studied in the literature: ***counterfactual responsibility*** [21] and ***responsibility for seeing to it*** [22]. To investigate these two notions and their properties, the rest of the paper is divided into six major sections:

- **Section 2** – With the help of a motivational example introduced in Subsection 2.1, we formally define and discuss the extensive form game model which captures the multi-step decision schemes in Subsection 2.2. Then, we give a discussion of the two forms of responsibility and related logic notions as well as a review of the

[2] For instance, young kids, animals, and autonomous agents are usually not treated as proper subjects of certain types of responsibility (*e.g.* legal responsibility). Section 5 contains a detailed discussion of the situation where an agent is improper to be held responsible.

corresponding literature in Subsection 2.3 and Subsection 2.4. In particular, a new form of seeing-to-it responsibility for extensive form game settings is proposed in Subsection 2.4.3.

- **Section 3** – The core terminology, the syntax of the modal language, and the semantics of the two forms of responsibility are formally defined in this section.
- **Section 4** – In this section, it is proved that the two forms of responsibility are not definable via each other in extensive form games. This result shows a type of independence between the two forms of responsibility and indicates the importance of including both of them into study.
- **Section 5** – The higher-order responsibility, which can be expressed with the nesting of the responsibility modalities, is discussed in this section. Its importance is significant in application scenarios where the responsible agent is not a *proper subject* to ascribe responsibility. Some properties of higher-order responsibility are proved, discussed, and compared with the results in the literature.
- **Section 6** – In this section, we formally study the *responsibility gap*, a situation where a statement is true in the outcome but no agent is responsible for it. Albeit the discussion of the responsibility gap is prevalent in the literature [8, 23–28], only a few studies [29, 30] give a formal definition of the concept. The *hierarchy* of responsibility gaps is then defined in Subsection 6.3, which, as far as we know, have never been discussed before. It is proved that a higher-order responsibility gap does not exist for sufficiently high orders.
- **Section 7** – This section investigates how the definitions of responsibility can be extended into imperfect information settings. It is found that the original definitions do not work properly in such settings and the responsibility gaps may always exist even when higher-order gaps are taken into account. After several failed attempts to modify the definition, it is observed that, if the seeing-to-it responsibility is defined based on the tree structure of the extensive form games, then no proper definition exists in imperfect information settings.

2 Responsibility and Decision Schemes

In this section, we introduce the formal model of the decision schemes that will be used throughout the rest of the paper. We also review the existing approaches to formally defining responsibility and discuss their shortcomings. To address some of these shortcomings, at the end of the section, we propose a new form of seeing-to-it responsibility.

2.1 Motivational Example

We start with a motivational example based on a real-life story. We use it as a running example throughout the whole paper. Some of the details of this example come into play only later in the paper.

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 [31]. 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 [32]. 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 [33, Article VII, Section 9]. With time, the law, the name of the board, and the way it grants approval changed [34], 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 Fig. 1a. First, the Board (agent b) can either deny (action D) the clemency or recommend (action E) it. If the Board recommends, then the Governor (agent g) might grant (action G) or not grant (action F) the executive clemency.

The executive clemency procedure in Montana is a typical multiagent sequential decision scheme where the final outcome is determined by the joint decision of all agents in the system. It is used in the rest of this paper as the running example to elucidate the notions and the claims.

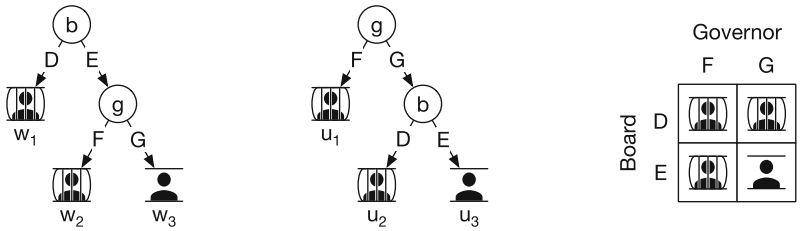
2.2 Decision Schemes

Throughout the paper, we 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

1. each non-leaf node is labelled with an agent;
2. each leaf node is labelled with a set of propositional variables.

Informally, each node of the tree represents a *state* of the game, while the root of the tree represents the initial state of this game. In each non-leaf node (*i.e.* non-terminal state), the agent labelling this node takes an action that chooses a child node as the



(a) Montana situation (b) Switched action order (c) Strategic form

Fig. 1 Executive clemency procedure

next state. The leaf nodes (*i.e.* terminal states) of a game are called *outcomes* of the game. An outcome is said to be labelled with a propositional variable if the outcome is labelled with a set containing this propositional variable, which intuitively means that the statement corresponding to the propositional variable is true in the outcome. The set of all outcomes of a game G is denoted by $\Omega(G)$. The notation $parent(n)$ refers to the parent node of any non-root node n. If node n_2 is on the simple path[3] (including ends) between the root node and node n_1, then write $n_1 \preceq n_2$.

As an example, consider the extensive form game in Fig. 1a and a propositional variable p which represents the statement "Beach is left in prison". The b-labelled node and the g-labelled node are non-leaf nodes while w_1, w_2, and w_3 are outcomes. The set $\Omega(G)$ in this game is $\{w_1, w_2, w_3\}$. Among all three outcomes, w_1 and w_2 are labelled with the set $\{p\}$ while w_3 is labelled with an empty set. At the b-labelled node (*i.e.* root), by recommending (E) the clemency, the Board chooses the g-labelled node as the next state. Then, by rejecting (F) the clemency, the Governor chooses w_2 as the next state, which is also the outcome. The path of play to outcome w_2 consists of three nodes: the b-labelled node, the g-labelled node, and outcome w_2. Among these three nodes, the b-labelled node (denoted by n_b here) is the parent node of the g-labelled node (denoted by n_g here), while the latter is the parent node of outcome w_2. This means $parent(n_g)$ is n_b and $parent(w_2)$ is n_g. Also, $w_2 \preceq n_b$ because the b-labelled node is on the simple path (as one of the ends) between the root node and outcome w_2.

In this paper, we use extensive form games as the models of multi-step decision schemes. Note that, to hold an agent responsible, the agent should have "free will" to make her choice (*a.k.a.* the control condition) and know how the decision affects the outcome (*a.k.a.* the epistemic condition) [20]. The former condition is captured by the tree structure of the extensive form games where, by item 1 of Definition 1, the agent labelling each non-leaf state has full control over which child would become the next state. To capture the latter condition, we assume the structure of the extensive form game to be common knowledge of all agents in the corresponding decision scheme. Considering that agents such as human beings usually have limited mentality [35], we further assume the tree structure to be finite to enable the epistemic condition above.

It is worth mentioning that the assumption of the tree structure implies the determinacy of the decision scheme. To capture the indeterminacy, one can add a dummy agent *Nature* into the system so that the nondeterministic transitions between states are captured by the decision of *Nature* [36]. However, our model cannot capture the probabilistic transitions. At the same time, the finite tree structure guarantees the termination of the decision process in an outcome. It also guarantees that the outcome uniquely specifies the path of play. In this sense, a property of the outcomes (as shown in item 2 of Definition 1) is indeed a property of the history.

We acknowledge that our model cannot capture the concurrent decisions of different agents. Neither does it capture the settings in which an agent does not know the previous decisions in the history. Such settings are called "imperfect information" settings [37]. We discuss them separately in Section 7. Our model is also not good for capturing long-term interactions between agents that do not have a natural starting point that can be viewed as a root node of the game tree.

[3] A simple path in a graph is a path without repeating nodes.

2.3 Counterfactual Responsibility

Counterfactual responsibility captures the *principle of alternative possibilities* [21, 38, 39]: *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 Fig. 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 had a strategy (*i.e.* 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 in outcome w_3. 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. This definition of counterfactual responsibility for extensive form games is introduced in [40]. It also appears in [41].

Note that, in order for Beach to be freed, both the Governor and the Board must support this decision. 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 Fig. 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 [10, 42–44], where all agents act concurrently and just once. In particular, the strategic forms of the two extensive form games in Figs. 1a and 1b are identical, as shown in Fig. 1c. As a result, no definition of counterfactual responsibility for strategic form settings can distinguish the two situations in Figs. 1a and 1b, which are different according to the above analysis.

2.4 Responsibility for Seeing To It

The other commonly studied form of responsibility is defined via the notion of *seeing-to-it*. As a modality, seeing-to-it has been well studied in STIT logic [45–48]. 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*[4] is usually required to exist to capture the intuition that no agent should be responsible

[4] In the general STIT models, a negative condition is a history where $\neg\varphi$ is true [49]. In the extensive form game settings, a negative condition is an *outcome* where $\neg\varphi$ is true. Note that a negative condition is distinct from the requirement of "an ability to prevent" in the definition of counterfactual responsibility. For example, by deciding not to compete in the Olympic games, an athlete sees to it that she does not win a gold Olympic medal. In this case, the negative condition is the outcome in which she wins the medal. This negative condition is potentially reachable if she does not give up. However, in this example, the athlete is not counterfactual responsible for not winning the medal because she has no strategy that guarantees winning the Olympics.

for a trivial truth[5] such as "the sun will rise". The notion of *deliberative* seeing-to-it [50–53] captures this idea by adding the requirement of a negative condition. Some follow-up work [54, 55] further incorporates the epistemic states of the agents into the discussion, but still within the STIT frame. Naumov and Tao [12] 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: *strategic* seeing-to-it in the presence of a negative condition and *achievement* seeing-to-it.

2.4.1 Strategic Seeing-To-It

Strategic seeing-to-it [47, 56, 57] is defined under the assumption that each agent commits upfront to a strategy 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 Fig. 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, strategic 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 strategic seeing-to-it has to be evaluated based on strategies rather than outcomes [58]. However, in some applications, such strategies may not be observable. Let us consider the case of outcome w_1 in Fig. 1a. Here, the strategy of the Governor is not observable because he has no chance to make any choice on the path of play to outcome w_1. 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 autonomous agents), the notion of strategic seeing-to-it accuses the agents of *mens rea* (*i.e.* guilty mind) purely based on their plans rather than actions. Note that, in law, *actus reus* (*i.e.* guilty action) is a commonly required element of a crime [59]. 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, *strategic seeing-to-it in the presence of a negative condition* cannot always serve as a proper notion of responsibility in extensive form games.

[5] A trivial truth is something that has been settled true within the criteria of concern. In an extensive form game, a trivial truth is a statement that is true in all outcomes.

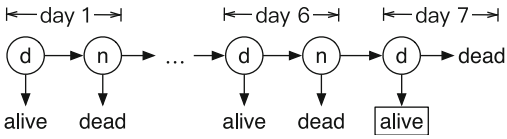


Fig. 2 An extensive form game between a doctor (d) and the *Nature* (n)

2.4.2 Achievement Seeing-To-It

Another notion of seeing-to-it that may capture a form of responsibility in extensive form game settings is *achievement seeing-to-it* [38, 50]. In an extensive form game, the agents make choices one after another[6]. 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 Fig. 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 Fig. 1a. When the executive clemency procedure is treated 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. If the patient is not given treatment, she will die within a week. Suppose the treatment is sure to cure the patient. We can model this scenario as a two-agent extensive form game between a doctor (d) and the *Nature* (n) shown in Fig. 2. Imagine the situation when 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 on any of the first six days. For this reason, *achievement seeing-to-it*

[6] Originally, achievement seeing-to-it has been proposed for STIT frames, where multiple agents can act simultaneously.

often cannot serve as a proper notion of the responsibility for seeing to it in extensive form games.

2.4.3 A New Notion of Responsibility for Seeing To It

As discussed above, neither strategic seeing-to-it in the presence of a negative condition nor achievement seeing-to-it can serve as a proper definition of seeing-to-it responsibility in extensive form games. In this subsection, we propose a new notion of seeing-to-it responsibility that fits into extensive form game settings.

First, we modify the notion of strategic seeing-to-it into a backward version. We would say that an agent *backwards-strategically* sees to φ if the agent has an upfront ability[7] 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 strategic 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 strategic seeing-to-it focuses on the *ability* of guaranteeing it.

Unlike the original notion of strategic 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 at the beginning of the game depicted in Fig. 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, we use the notion of backwards-strategic seeing-to-it, in combination with achievement seeing-to-it, to define the seeing-to-it form of responsibility in extensive form game settings. We 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, we 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 Fig. 1a. In outcome w_1, the Board sees to Beach being left in prison both backwards-strategically and in the achievement way. Therefore, the

[7] By an "ability" here and in the rest of the paper, we mean a *strategic* ability – having a strategy that guarantees certain statement to be true no matter what are the actions of the other agents.

Board is responsible for seeing to Beach being left in prison in outcome w_1. This argument is also true for the Governor in outcome w_2. However, in outcome 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 outcome w_1. In outcome 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 outcome 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.

Similarly, in our "doctor and *Nature*" example from Fig. 2, the doctor had a strategy to cure the patient in the first state (*i.e.* giving a treatment). But she did not maintain it when the game transitioned to the second state, where the *Nature* could let the patient die. This means the doctor does not see to the patient being cured backwards-strategically. Thus, the doctor is not responsible for seeing to the patient being cured.

3 Terminology, Syntax, and Semantics

We first define two terms based on the tree structure of extensive form games.

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 the set X is 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 set X. For example, in the extensive form game depicted in Fig. 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.

Note that, unless $X = \Omega(G)$, a negative condition must be available at the root node. However, when an outcome $w \in X$ is reached, no negative condition is available. In between (*i.e.* on the path from the root to outcome w), there must be a *unique* moment when the last possibility of a negative condition is eliminated. For example, in the extensive form game in Fig. 1a, the Board denying (D) the clemency is the unique moment when the last possibility to set Beach free is eliminated on the path of play to outcome w_1, the Governor rejecting (F) the clemency is such a unique moment on the path of play to outcome w_2, and the Governor granting (G) the clemency is the

unique moment when the last possibility to leave Beach in prison is eliminated on the path of play to outcome w_3. Observe that, by Definition 2, the uniqueness of such a moment implies the uniqueness of an achievement point on the path of play, which is formally captured in the next lemma.

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 $w \preceq n$.*

The term "achievement point" in Definition 2 is used to capture the notion of achievement seeing-to-it. Next, let us consider a notation, $win_a(X)$, that will be used to capture the notion of backwards-strategic seeing-to-it. For any set X of outcomes and any agent a, by $win_a(X)$, we mean the set of all nodes (including outcomes) from which agent a has the *ability* to end the game in set X. Informally, in an outcome w, such ability exists if and only if $w \in X$; in a non-leaf node, it is captured by the existence of a strategy to achieve an outcome in set X. Inspired by the *minimax algorithm* in zero-sum games [60], the set $win_a(X)$ is formally defined below using backward induction.

Definition 3 For any set X of outcomes, the set $\boldsymbol{win_a(X)}$ 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 $n \in win_a(X)$;
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 $n \in win_a(X)$.

In particular, 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 $win_a(X)$. Moreover, 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. In the game in Fig. 1a, by Definition 3, for the set $X = \{w_1, w_2\}$ where Beach is left in prison, the set $win_b(X)$ consists of outcome w_1, outcome w_2, and the b-labelled node, while the set $win_g(X)$ consists of outcome w_1, outcome w_2, the g-labelled node, and the b-labelled node. This is consistent with the analysis in Subsection 2.4.3.

To formally study the counterfactual responsibility and the seeing-to-it responsibility, in this paper, we use the modal language Φ 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, $a \in \mathcal{A}$ is an agent, and modalities C and S are used to express the counterfactual responsibility and the seeing-to-it responsibility, respectively. In particular, 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". In this modal language, Boolean constants true $\top$ and false $\bot$ are defined in the standard way.

The next definition is at the core 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 is 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 [40, 41]. 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-strategic 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 $win_a([\![\varphi]\!])$, as part 5a 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 $[\![\varphi]\!]$-achievement point by agent a. This is captured in part 5b of Definition 4.

4 Mutual Undefinability of Modalities C and S

In the previous sections, we introduced two notions of responsibility: counterfactual and seeing to it. Let us now study a possible connection between these two notions. Potentially, they can be connected in many different ways. First, they can be *equivalent*: an agent is responsible for seeing to it if and only if she is responsible counterfactually. Second, one of these two notions can be a *special case* of the other. If this were the case, then each time when an agent is responsible under the more special notion or responsibility, she would also be responsible under the more general notion of responsibility. However, as we have seen, in outcome w_3 of the extensive form game captured in Fig. 1a, the Governor is counterfactually responsible for Beach being set free but not responsible for seeing to this; in outcome w_1 of the same game, the Board is responsible for seeing to Beach being left in prison but not responsible for this counterfactually. Thus, there exists neither the first nor the second type of connection between the two concepts of responsibility.

Another possible connection between notions is *definability* of one of them via the other. For example, some people argue that knowledge is a justified true belief [61].

Those who share this view think that the concept of knowledge is definable (or expressible) via the concepts of justification, truth, and belief. In this section, we will prove that a similar definability connection does not exist[8] for the two forms of responsibility. Proving that a definability connection exists is relatively simple: one just needs to state the connection explicitly (i.e. "knowledge is a justified true belief") and give an argument in support of it. Proving that such a connection does not exist requires giving a counterexample for *each* potential way to express one notion via another. Before giving such proof, one needs to decide on the language in which the connection will be considered. In this paper, we show that counterfactual and seeing-to-it forms of responsibility are not definable via each other in the modal language Φ.

Note that our results do not imply that there is absolutely no connection between these two notions of responsibility. In Subsection 5.3, we give an example of a formula with nested modalities which is valid in all extensive form games. The existence of such formulae demonstrates that there is some, albeit a very weak, connection between the two forms of responsibility.

Our undefinability results are formally stated in Theorems 1 and 2. Towards the proofs, let us first introduce an auxiliary definition:

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

We would say that, 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 modality S via modality C could be specified similarly.

To prove the undefinability results, we use a technique named "truth set algebra", which is introduced in [63] and also used in [64]. Unlike the traditional "bisimulation" method, the "truth sets algebra" technique uses a single model. Generally speaking, to prove the undefinability of modality C via modality S, a specific extensive form game is defined and used to show the semantic inequivalence between formula $\mathsf{C}_a p$ and any formula in language Φ that does not use modality C. The proof of the undefinability of modality S via modality C is similar.

4.1 Undefinability of Modality C via Modality S

Throughout this subsection, consider an extensive form game between agents a, b, and c depicted at the top of Fig. 3. It has four outcomes: w_1, w_2, w_3, and w_4. Without loss of generality[9], in this subsection, assume that language Φ contains only agents a, b, c and a single propositional variable p. Outcomes w_1 and w_3 are labelled with the set $\{p\}$ and outcomes w_2 and w_4 are labelled with the empty set. In the middle

[8] Such connections might exist or not for similar notions in other settings. For example, Naumov and Tao [12] showed that counterfactual responsibility is definable via a *deliberative* STIT sense of responsibility but not vice versa in a single-step (strategic) game setting. At the same time, Cui and Naumov [62] observed that counterfactual responsibility is not definable via an *achievement* STIT sense of responsibility and vice versa in extensive form games with trees of infinite depth.

[9] Alternatively, additional agents and propositional variables can be assumed to be present but not used as labels. In particular, according to items 4 and 5 of Definition 4, it can be deduced that $[\![\mathsf{S}_d\varphi]\!] = [\![\mathsf{C}_d\varphi]\!] = \varnothing = [\![\bot]\!]$ for any agent d which is *not* used as a label in the game and any formula $\varphi \in \Phi$.

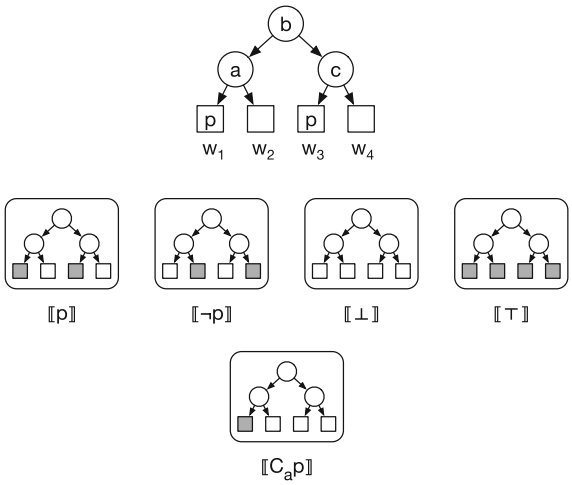


Fig. 3 Towards the proof of undefinability of C via S

part of Fig. 3, four miniaturised game trees are used to visualise the truth sets $[\![p]\!]$, $[\![\neg p]\!]$, $[\![\bot]\!]$, and $[\![\top]\!]$. Specifically, the truth set $[\![p]\!]$ is visualised by *shading grey* the outcomes in the miniaturised tree that belong to the set $[\![p]\!]$. The same is true for the other three truth sets. Denote by $\mathcal{F}$ the family $\{[\![p]\!], [\![\neg p]\!], [\![\bot]\!], [\![\top]\!]\}$ of truth sets.

To prove the undefinability of modality C via modality S, by Definition 5, it suffices to show that, in the game depicted at the top of Fig. 3, the truth set $[\![\mathsf{C}_a p]\!]$ is not equal to the truth set $[\![\varphi]\!]$ for each formula $\varphi \in \Phi$ that does not use modality C. We will prove this by showing that, for each such formula φ, the truth set $[\![\varphi]\!]$ is a member of the family $\mathcal{F}$, while the truth set $[\![\mathsf{C}_a p]\!]$, as depicted at the bottom of Fig. 3, is not in the family $\mathcal{F}$. The formal proof is given below.

Lemma 2 *For any formulae $\varphi, \psi \in \Phi$, if $[\![\varphi]\!], [\![\psi]\!] \in \mathcal{F}$, then $[\![\neg\varphi]\!], [\![\varphi \wedge \psi]\!] \in \mathcal{F}$.*

Proof Observe that, the family $\mathcal{F}$ is closed with respect to complement and intersection. Then, the statement of this lemma follows from items 2 and 3 of Definition 4. □

Lemma 3 $[\![\mathsf{S}_g\varphi]\!] = [\![\bot]\!]$ *for each agent $g \in \{a, b, c\}$ and each formula $\varphi \in \Phi$ such that $[\![\varphi]\!] \in \mathcal{F}$.*

Proof We first show that if $[\![\varphi]\!] = [\![p]\!]$, then $[\![\mathsf{S}_g\varphi]\!] = [\![\bot]\!]$ for each agent $g \in \{a, b, c\}$. Indeed, $[\![\varphi]\!] = [\![p]\!] = \{w_1, w_3\}$. Then, by Definition 3, the root node of the tree does *not* belong to the set $win_g([\![\varphi]\!])$ for each agent $g \in \{a, b, c\}$. Thus, for each agent $g \in \{a, b, c\}$, there is no single path from the root to an outcome such that all nodes of this path belong to the set $win_g([\![\varphi]\!])$. Hence, by item 5a of Definition 4, none of the agents is responsible for seeing to φ in any of the outcomes. Therefore, for each formula $\varphi \in \Phi$ and each agent $g \in \{a, b, c\}$, if $[\![\varphi]\!] = [\![p]\!]$, then $[\![\mathsf{S}_g\varphi]\!] = [\![\bot]\!]$.

The justification for the cases where $[\![\varphi]\!] = [\![\neg p]\!]$ is similar to the above case.

In the case where $[\![\varphi]\!] = [\![\bot]\!]$, note that the set $[\![\bot]\!]$ is empty. Thus, $[\![\varphi]\!] = \varnothing$. Then, $win_g([\![\varphi]\!]) = \varnothing$ for each agent $g \in \{a, b, c\}$ by Definition 3. Hence, by item 5a of Definition 4, none of the agents is responsible for seeing to φ in any of the outcomes.

Therefore, for each formula $\varphi \in \Phi$ and each agent $g \in \{a, b, c\}$, if $[\![\varphi]\!] = [\![\bot]\!]$, then $[\![\mathsf{S}_g\varphi]\!] = [\![\bot]\!]$.

In the case where $[\![\varphi]\!] = [\![\top]\!]$, observe that the set $[\![\top]\!]$ is the set of all outcomes in the game. Thus, there is no $[\![\varphi]\!]$-achievement point by Definition 2. Hence, by item 5b of Definition 4, none of the agents is responsible for seeing to φ in any of the outcomes. Therefore, for each formula $\varphi \in \Phi$ and each agent $g \in \{a, b, c\}$, if $[\![\varphi]\!] = [\![\top]\!]$, then $[\![\mathsf{S}_g\varphi]\!] = [\![\bot]\!]$. □

Lemma 4 $[\![\varphi]\!] \in \mathcal{F}$ *for any formula* φ *that does not use modality* C.

Proof We prove the statement of the lemma by induction on the structural complexity of formula φ.

If φ is propositional variable p, then the statement of the lemma is true because the truth set $[\![p]\!]$ is an element of the family $\mathcal{F}$.

If formula φ has the form $\neg\psi$ or $\psi_1 \wedge \psi_2$, then $[\![\psi]\!] \in \mathcal{F}$ or $[\![\psi_1]\!], [\![\psi_2]\!] \in \mathcal{F}$ by the induction hypothesis. In this case, the statement of the lemma follows from Lemma 2.

If formula φ has the form $\mathsf{S}_g\psi$, where $g \in \{a, b, c\}$, then $[\![\psi]\!] \in \mathcal{F}$ by the induction hypothesis. In this case, the statement of the lemma follows from Lemma 3 and that $[\![\bot]\!] \in \mathcal{F}$. □

Lemma 5 $[\![\mathsf{C}_a\, p]\!] \notin \mathcal{F}$.

Proof Indeed, $[\![\neg p]\!] = \{w_2, w_4\}$. Then, the node labelled with agent a belongs to the set $win_a([\![\neg p]\!])$ by Definition 3. This means agent a has a strategy ("go right") to prevent p on the path to outcome w_1. However, agent a has no such strategy on the path to outcome w_3. Hence, $[\![\mathsf{C}_a\, p]\!] = \{w_1\}$ by item 4 of Definition 4. Therefore, $[\![\mathsf{C}_a\, p]\!] \notin \mathcal{F}$. □

The next theorem follows from Definition 5 and the two previous lemmas.

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 use modality* C.

Intuitively, this result shows that the notion of individual counterfactual responsibility for an arbitrary fact p cannot be defined (expressed) via our form of seeing-to-it responsibility even in a very complicated way.

4.2 Undefinability of Modality S via Modality C

By Definition 5, to show that a formula φ is not semantically equivalent to any formula in a language Ψ, it suffices for each formula $\psi \in \Psi$ to construct a game (model) G_ψ such that $[\![\varphi]\!] \neq [\![\psi]\!]$ in game G_ψ. In Subsection 4.1, we managed to construct a *uniform* game that does not depend on formula ψ. This made the whole proof relatively simple. However, it is not clear how to construct such a uniform game to prove the undefinability result in this subsection. As a compromise, the proof here constructs a different game G_ψ for each formula $\psi \in \Psi$. More precisely, for each formula $\psi \in \Phi$ that does not use modality S, a game G_N where N is the number of occurrences of modality C in formula ψ is constructed. We formally describe the construction below.

Without loss of generality, in this subsection, assume that language Φ has a single propositional variable p and two agents, a and b. Consider a game G_N with a parameter N, where N is an *arbitrary* positive integer. Game G_N is depicted at the top of Fig. 4. It has $N + 3$ non-leaf nodes and $N + 4$ outcomes (leaf nodes): w_1, ..., w_{N+4}. The non-leaf nodes are labelled with agents a and b, who take turns to decide whether to terminate the game (by going down) or to continue (by going to the right). The game terminates after at most $N + 3$ turns. Which agent makes the last move depends on the parity of N. Outcomes w_1 and w_{N+4} are labelled with the empty set while outcomes w_2, ..., w_{N+3} are labelled with the set $\{p\}$.

For each integer n such that $3 \leq n \leq N + 3$, consider four *families* of truth sets: α_n, β_n, γ_n, and δ_n. The family α_n of truth sets consists of *all* subsets of the set $\{w_1, \ldots, w_{N+4}\}$ that exclude outcome w_1 and include outcomes $w_2, \ldots, w_n$:

$$\alpha_n := \big\{\{w_2, \ldots, w_n\} \cup X \mid X \subseteq \{w_{n+1}, \ldots, w_{N+4}\}\big\}.$$

The other families of truth sets are similarly defined as:

$$\beta_n := \big\{X \mid X \subseteq \{w_{n+1}, \ldots, w_{N+4}\}\big\};$$
$$\gamma_n := \big\{\{w_1\} \cup X \mid X \subseteq \{w_{n+1}, \ldots, w_{N+4}\}\big\};$$
$$\delta_n := \big\{\{w_1, w_2, \ldots, w_n\} \cup X \mid X \subseteq \{w_{n+1}, \ldots, w_{N+4}\}\big\}.$$

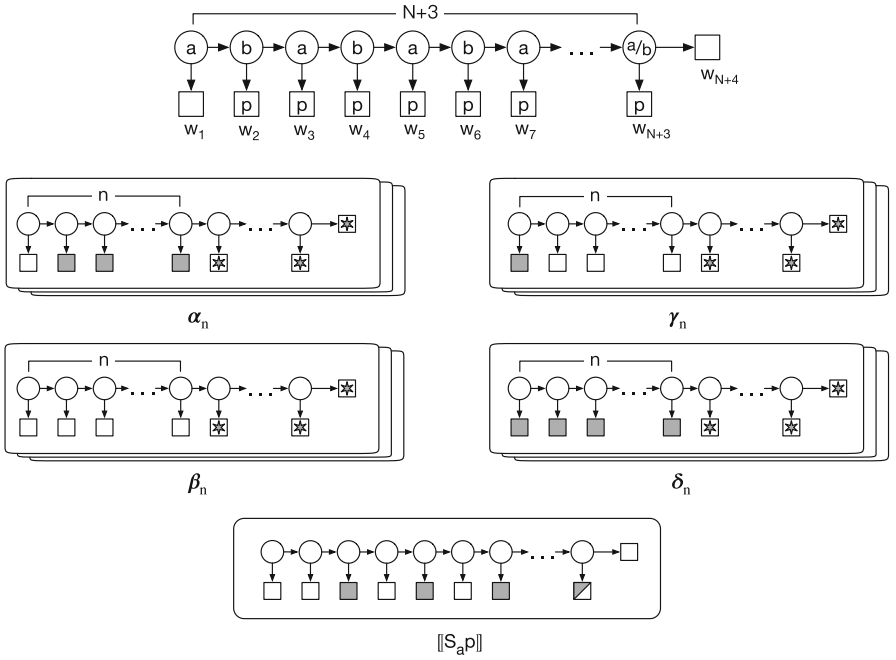


Fig. 4 Towards the proof of undefinability of S via C

The families α_n, β_n, γ_n, and δ_n are visualised in the middle two rows of Fig. 4. In these miniaturised game trees, the asterisk $*$ is used as the *wildcard* to mark the outcomes that *might but do not have to* belong to a set in the corresponding family.

Lemma 6 *For any formulae $\varphi, \psi \in \Phi$ and any $n \geq 0$, if $[\![\varphi]\!], [\![\psi]\!] \in \alpha_n \cup \beta_n \cup \gamma_n \cup \delta_n$, then $[\![\neg\varphi]\!], [\![\varphi \wedge \psi]\!] \in \alpha_n \cup \beta_n \cup \gamma_n \cup \delta_n$.*

Proof Observe that, by the definition of families α_n, β_n, γ_n, and δ_n, the family of sets $\alpha_n \cup \beta_n \cup \gamma_n \cup \delta_n$ is closed with respect to complement and intersection. Then, the statement of this lemma follows from items 2 and 3 of Definition 4. □

Lemma 7 *For any integer $n \geq 3$ and any formula $\varphi \in \Phi$,*

1. *if $[\![\varphi]\!] \in \alpha_n$, then $[\![\mathsf{C}_a\varphi]\!] \in \alpha_n$ and $[\![\mathsf{C}_b\varphi]\!] \in \beta_{n-1}$;*
2. *if $[\![\varphi]\!] \in \beta_n$, then $[\![\mathsf{C}_a\varphi]\!] \in \beta_n$ and $[\![\mathsf{C}_b\varphi]\!] \in \beta_n$;*
3. *if $[\![\varphi]\!] \in \gamma_n$, then $[\![\mathsf{C}_a\varphi]\!] \in \gamma_n$ and $[\![\mathsf{C}_b\varphi]\!] \in \beta_n$;*
4. *if $[\![\varphi]\!] \in \delta_n$, then $[\![\mathsf{C}_a\varphi]\!] \in \beta_{n-1}$ and $[\![\mathsf{C}_b\varphi]\!] \in \beta_{n-1}$.*

Proof Suppose that $[\![\varphi]\!] \in \alpha_n$ for some integer $n \geq 3$. Then,

$$w_1 \notin [\![\varphi]\!] \tag{1}$$

and

$$w_2, \ldots, w_n \in [\![\varphi]\!]. \tag{2}$$

Eq. 1 implies

$$w_1 \notin [\![\mathsf{C}_a\varphi]\!] \tag{3}$$

by item 4 of Definition 4 and $w_1 \in [\![\neg\varphi]\!]$ by item 2 of Definition 4. Then, the root node, which is labelled by agent a, belongs to the set $win_a([\![\neg\varphi]\!])$ by Definition 3. Hence,

$$w_2, \ldots, w_n \in [\![\mathsf{C}_a\varphi]\!] \tag{4}$$

by Eq. 2 and item 4 of Definition 4. Therefore, $[\![\mathsf{C}_a\varphi]\!] \in \alpha_n$ by Eqs. 3 and 4 and the definition of family α_n.

Note that, by Eq. 1 and item 4 of Definition 4,

$$w_1 \notin [\![\mathsf{C}_b\varphi]\!]. \tag{5}$$

Also, observe that none of the non-leaf nodes above outcomes $w_1, \ldots, w_{n-1}$ belongs to the set $win_b([\![\neg\varphi]\!])$ by Definition 3 because of Eqs. 1 and 2 and that the root node above outcome w_1 is labelled by agent a. Hence,

$$w_2, \ldots, w_{n-1} \notin [\![\mathsf{C}_b\varphi]\!] \tag{6}$$

by item 4 of Definition 4. Therefore, $[\![\mathsf{C}_b\varphi]\!] \in \beta_{n-1}$ by Eqs. 5 and 6 and the definition of family β_n.

The proofs of the other three parts of the lemma are similar. □

Lemma 8 *For any formula $\varphi \in \Phi$, any agent $g \in \{a, b\}$, and any integer $n \geq 3$, if $[\![\varphi]\!] \in \alpha_n \cup \beta_n \cup \gamma_n \cup \delta_n$, then $[\![\mathsf{C}_g\varphi]\!] \in \alpha_{n-1} \cup \beta_{n-1} \cup \gamma_{n-1} \cup \delta_{n-1}$.*

Proof Note that, $\alpha_n \subseteq \alpha_{n-1}$, $\beta_n \subseteq \beta_{n-1}$, $\gamma_n \subseteq \gamma_{n-1}$, and $\delta_n \subseteq \delta_{n-1}$ for each $n \geq 1$. Then, the statement of this lemma follows from Lemma 7. □

Lemma 9 *For any formula φ that does not use modality S and any positive integer $k \leq N$, if formula φ contains at most k occurrences of modality C, then,*

$$[\![\varphi]\!] \in \alpha_{N+3-k} \cup \beta_{N+3-k} \cup \gamma_{N+3-k} \cup \delta_{N+3-k}.$$

Proof We prove the statement of the lemma by induction on the structural complexity of formula φ.

If φ is propositional variable p, then $[\![\varphi]\!] = \{w_2, \ldots, w_{N+3}\} \in \alpha_{N+3} \subseteq \alpha_{N+3-k}$.

If formula φ is a disjunction or a negation, then the statement of the lemma follows from the induction hypothesis by Lemma 6.

If formula φ has the form $\mathsf{C}_g\psi$, where $g \in \{a, b\}$, then formula ψ contains at most $k-1$ occurrences of modality C. Thus, $[\![\varphi]\!] \in \alpha_{N+4-k} \cup \beta_{N+4-k} \cup \gamma_{N+4-k} \cup \delta_{N+4-k}$ by the induction hypothesis. Then, the statement of the lemma follows from Lemma 8. □

The truth set $[\![\mathsf{S}_a p]\!]$ is shown at the bottom of Fig. 4. However, to prove the undefinability result, it is enough to observe the following lemma:

Lemma 10 *$w_2 \notin [\![\mathsf{S}_a p]\!]$ and $w_3 \in [\![\mathsf{S}_a p]\!]$.*

Proof Observe that, in game G_N depicted at the top of Fig. 4, node w_2 is the $[\![p]\!]$-achievement point by agent b on the path of play toward outcome w_2. Then, by Lemma 1, there is no $[\![p]\!]$-achievement point by agent a on the path of play toward outcome w_2. Hence, $w_2 \notin [\![\mathsf{S}_a p]\!]$ by item 5b of Definition 4.

The non-leaf nodes above outcomes w_1, w_2, w_3 belong to the set $win_a([\![p]\!])$ by Definition 3. At the same time, node w_3 is the $[\![p]\!]$-achievement point by agent a on the path of play toward outcome w_3. Hence, $w_3 \in [\![\mathsf{S}_a p]\!]$ by item 5 of Definition 4. □

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 use modality S.*

Proof Assume the opposite. Then, by Definition 5, there is a formula $\varphi \in \Phi$ not using modality S such that $[\![\mathsf{S}_a p]\!] = [\![\varphi]\!]$ in every extensive form game. Let N be the number of occurrences of modality C in formula φ. Then, by Lemma 9, in game G_N,

$$\begin{aligned}[\![\mathsf{S}_a p]\!] = [\![\varphi]\!] &\in \alpha_{N+3-N} \cup \beta_{N+3-N} \cup \gamma_{N+3-N} \cup \delta_{N+3-N} \\ &= \alpha_3 \cup \beta_3 \cup \gamma_3 \cup \delta_3.\end{aligned}$$

However, $[\![\mathsf{S}_a p]\!] \notin \alpha_3 \cup \beta_3 \cup \gamma_3 \cup \delta_3$ by Lemma 10 and the definition of families α_3, β_3, γ_3, and δ_3. □

Theorems 1 and 2 above 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.

5 Higher-Order Responsibility

Intuitively, it seems natural to think that an agent can be responsible for the fact that another person is responsible. For instance, if a criminal escapes from prison due to the negligence of a security guard and murders someone, then the escaped prisoner is responsible for the murder but the guard is responsible for the prisoner being responsible. In this paper, we call such responsibility (*e.g.* the responsibility of the guard) for another agent's responsibility a "higher-order responsibility".

The 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 [65]. The same is true in many other situations such as animals and their owners, autonomous machines and their designers, as well as automatic weapons and their commanders. For example, when discussing the responsibility issue of automatic weapons, Himmelreich [66] identified a situation that "a merely minimal agent does φ such that no one (*i.e.* human person) is responsible for φ; but had φ been the action of a human person, then this person would be responsible for φ". In this situation, a human person is treated as a proper subject of responsibility while a "merely minimal agent" is not. Hindriks and Veluwenkamp [67] rephrased this situation in the context of AI as "an autonomous machine causes harm, no one is to blame for it, but the blame would be appropriate had it been caused by a human being". To tackle this situation, they further came up with the notion of "indirect responsibility", which can be captured by the higher-order responsibility discussed here. In this way, the discussion of higher-order responsibility offers a new perspective on dealing with the responsibility issue of automatic weapons, which has raised many concerns [68–70].

The existing notions of Chellas [45]'s seeing-to-it, deliberative seeing-to-it, and achievement seeing-to-it cannot capture higher-order responsibility because the nesting of two corresponding modalities trivialises somehow. For example, if modality $\Box$ represents Chellas [45]'s seeing-to-it, then, for distinct agents a and b, the statement $\Box_a\Box_b\varphi$ is equivalent to φ being unavoidably true. Nesting counterfactual responsibility in a one-step (strategic) game setting [43] also does not meaningfully capture the higher-order responsibility. As we discuss below, in the extensive form game setting, our version of seeing-to-it responsibility modality and the counterfactual responsibility modality do not always trivialise and are capable of capturing different forms of higher-order responsibility.

5.1 Idempotency of Responsibility

Perhaps one of the most unexpected properties of the proposed seeing-to-it modality S is the lack of idempotency[10]. That is, formula $\mathsf{S}_a\mathsf{S}_a\varphi$ is not, generally speaking,

[10] Idempotency is a property of some operations that can be repeated without changing the result beyond the initial application.

semantically equivalent to $\mathsf{S}_a\varphi$. In contrast, Chellas [45]'s STIT, deliberative STIT, and achievement STIT modalities are all idempotent [50]. The fact that our modality is not idempotent can be observed in outcome w_2 of the game depicted in Fig. 1a. Consider the Governor g and formula φ which represents "Beach is left in prison". Then, $w_2 \in [\![\mathsf{S}_g\varphi]\!]$ and $w_1 \notin [\![\mathsf{S}_g\varphi]\!]$. By item 3 of Definition 3, the root node (*i.e.* the b-labelled node) does not belong to the set $win_g([\![\mathsf{S}_g\varphi]\!])$. Thus, $w_2 \notin [\![\mathsf{S}_g\mathsf{S}_g\varphi]\!]$ by item 5a of Definition 4. In other words, even though the Governor is responsible for seeing to it that Beach is left in prison in outcome w_2, he is not responsible for seeing to himself assuming this responsibility. Therefore, $[\![\mathsf{S}_g\mathsf{S}_g\varphi]\!] \neq [\![\mathsf{S}_g\varphi]\!]$ in the game depicted in Fig. 1a.

At the same time, as we show below, counterfactual responsibility modality C_a, as defined in item 4 of Definition 4, is idempotent. The same is also true in the strategic game setting [43]. To show that "if an agent is counterfactually responsible for a statement φ, then she is also counterfactually responsible for assuming this counterfactual responsibility", it is enough to prove that $[\![\mathsf{C}_a\varphi]\!] \subseteq [\![\mathsf{C}_a\mathsf{C}_a\varphi]\!]$ for any agent a, any formula $\varphi \in \Phi$, and any extensive form game. Here, we claim a stronger statement as stated in Proposition 1 below. To prove it, we first show the next lemma.

Lemma 11 *For any formulae $\varphi, \psi \in \Phi$ and any agent a, if $[\![\varphi]\!] \subseteq [\![\psi]\!]$, then $win_a([\![\varphi]\!]) \subseteq win_a([\![\psi]\!])$.*

Proof We prove this lemma by backward induction in the game tree, from the leaf nodes up to the root. We show that, for each node n in the tree, if $n \in win_a([\![\varphi]\!])$, then $n \in win_a([\![\psi]\!])$.

If node n is a leaf node, then the assumption $n \in win_a([\![\varphi]\!])$ implies $n \in [\![\varphi]\!]$ by item 1 and the requirement of "minimal set" in Definition 3. Then, $n \in [\![\psi]\!]$ by the assumption $[\![\varphi]\!] \subseteq [\![\psi]\!]$ of the lemma. Therefore, $n \in win_a[\![\psi]\!]$ by item 1 of Definition 3.

If node n is a non-leaf node, then n is labelled by either agent a or some other agent. In the first case, by item 2 of Definition 3, the assumption $n \in win_a([\![\varphi]\!])$ implies the existence of n's child node m such that $m \in win_a([\![\varphi]\!])$. Then, $m \in win_a([\![\psi]\!])$ by the induction hypothesis. Therefore, $n \in win_a([\![\psi]\!])$ by item 2 of Definition 3.

In the second case, by item 3 of Definition 3, the assumption $n \in win_a([\![\varphi]\!])$ implies that every child node m of node n satisfies that $m \in win_a([\![\varphi]\!])$. Then, by the induction hypothesis, $m \in win_a([\![\psi]\!])$. Therefore, $n \in win_a([\![\psi]\!])$ by item 3 of Definition 3. □

Proposition 1 *For any formula $\varphi \in \Phi$ and any agent $a \in \mathcal{A}$, formulae $\mathsf{C}_a\mathsf{C}_a\varphi$ and $\mathsf{C}_a\varphi$ are semantically equivalent.*

Proof By Definition 5, it suffices to show that for any outcome w in any extensive form game, $w \in [\![\mathsf{C}_a\mathsf{C}_a\varphi]\!]$ if and only if $w \in [\![\mathsf{C}_a\varphi]\!]$.

For the "*if*" part, consider an arbitrary outcome $w \in [\![\mathsf{C}_a\varphi]\!]$. By item 4 of Definition 4, there is a node n such that $w \preceq n$ and

$$n \in win_a([\![\neg\varphi]\!]). \tag{7}$$

At the same time, $[\![\varphi]\!] \supseteq [\![\mathsf{C}_a\varphi]\!]$ by item 4 of Definition 4. Then, $[\![\neg\varphi]\!] \subseteq [\![\neg\mathsf{C}_a\varphi]\!]$ by item 2 of Definition 4. It further implies that $win_a([\![\neg\varphi]\!]) \subseteq win_a([\![\neg\mathsf{C}_a\varphi]\!])$ by

Lemma 11. Hence, $n \in win_a([\![\neg \mathsf{C}_a\varphi]\!])$ by Eq. 7. Together with the statement that $w \preceq n$ and the assumption $w \in [\![\mathsf{C}_a\varphi]\!]$, it can be concluded that $w \in [\![\mathsf{C}_a\mathsf{C}_a\varphi]\!]$ by item 4 of Definition 4.

The "*only if*" part follows directly from item 4 of Definition 4. □

5.2 Single-Form Higher-Order Responsibility

In the previous subsection, we observed that nesting modality C for *the same* agent trivialises while nesting S with the same agent does not. In this subsection, we observe that nesting of the same modalities with *different* agents behaves in the opposite way: $\mathsf{S}_a\mathsf{S}_b$ trivialises, while $\mathsf{C}_a\mathsf{C}_b$ does not. To show the former, in Proposition 2 we show the semantic equivalence of formulae $\mathsf{S}_b\mathsf{S}_a\varphi$ and $\bot$ for distinct agents a and b. To increase the readability of its proof, we first show the next two lemmas.

Lemma 12 *$[\![\mathsf{S}_a\varphi]\!] \subseteq [\![\varphi]\!]$ for any formula $\varphi \in \Phi$, any agent $a \in \mathcal{A}$, and any extensive form game G.*

Proof Note that $[\![\mathsf{S}_a\varphi]\!] \subseteq \Omega(G)$ and $[\![\mathsf{S}_a\varphi]\!] \subseteq win_a([\![\varphi]\!])$ by item 5 of Definition 4. Then,

$$[\![\mathsf{S}_a\varphi]\!] \subseteq \Omega(G) \cap win_a([\![\varphi]\!]). \tag{8}$$

Meanwhile, $\Omega(G) \cap win_a([\![\varphi]\!]) = [\![\varphi]\!]$ by item 1 and the minimality condition in Definition 3. Hence, $[\![\mathsf{S}_a\varphi]\!] \subseteq [\![\varphi]\!]$ by Eq. 8.

Lemma 13 *For any outcomes w and w', any formula φ, and any $[\![\varphi]\!]$-achievement point n by some agent, if $w \preceq n$, $w' \preceq n$, and $w \in [\![\mathsf{S}_a\varphi]\!]$, then $w' \in [\![\mathsf{S}_a\varphi]\!]$.*

Proof By Lemma 1 and the assumption of the current lemma, node n is the unique $[\![\varphi]\!]$-achievement point on the path of play to outcome w. Then, node n is the $[\![\varphi]\!]$-achievement point by agent a according to item 5b of Definition 4 and the assumption $w \in [\![\mathsf{S}_a\varphi]\!]$ of the lemma. By item 3 of Definition 2, the statement that node n is the $[\![\varphi]\!]$-achievement point implies that $w'' \in [\![\varphi]\!]$ for each outcome w'' such that $w'' \preceq n$. Thus, by Definition 3,

$$\{m \mid m \preceq n\} \subseteq win_a([\![\varphi]\!]). \tag{9}$$

By the assumption $w \preceq n$ of the lemma, the set of nodes on the path from the root to node n is a subset of the set of nodes on the path from the root to outcome w. That is,

$$\{m' \mid n \preceq m'\} \subseteq \{m' \mid w \preceq m'\}. \tag{10}$$

On the other hand, by item 5a of Definition 4, the assumption $w \in [\![\mathsf{S}_a\varphi]\!]$ of the lemma implies that $\{m' \mid w \preceq m'\} \subseteq win_a([\![\varphi]\!])$. Hence, by Eq. 10,

$$\{m' \mid n \preceq m'\} \subseteq win_a([\![\varphi]\!]). \tag{11}$$

Observe that, by the assumption $w' \preceq n$ of the lemma, every node on the path from the root to outcome w' is in the set $\{m \mid m \preceq n\}$ or in the set $\{m' \mid n \preceq m'\}$. Thus, by Eqs. 9 and 11,

$$\{m'' \mid w' \preceq m''\} \subseteq win_a([\![\varphi]\!]). \tag{12}$$

Note that, node n is also the $[\![\varphi]\!]$-achievement point by agent a on the path of play to outcome w'. Therefore, $w' \in [\![\mathsf{S}_a\varphi]\!]$ by Eq. 12 and item 5 of Definition 4. □

Proposition 2 *For any formula $\varphi \in \Phi$ and any distinct agents $a, b \in \mathcal{A}$, formula $\mathsf{S}_b\mathsf{S}_a\varphi$ is semantically equivalent to $\bot$.*

Proof By Definition 5, it suffices to prove $[\![\mathsf{S}_b\mathsf{S}_a\varphi]\!] = \varnothing$ for each extensive form game. Moreover, $[\![\mathsf{S}_b\mathsf{S}_a\varphi]\!] \subseteq [\![\mathsf{S}_a\varphi]\!]$ by Lemma 12. Thus, it suffices to show that $w \notin [\![\mathsf{S}_b\mathsf{S}_a\varphi]\!]$ for each outcome $w \in [\![\mathsf{S}_a\varphi]\!]$.

Consider an outcome $w \in [\![\mathsf{S}_a\varphi]\!]$. Then, by item 5b of Definition 4 and Definition 2, there is a $[\![\varphi]\!]$-achievement point n by agent a such that

1. $w \preceq n$;
2. $parent(n)$ is labelled with agent a;
3. there exists an outcome w' such that $w' \preceq parent(n)$ and $w' \notin [\![\varphi]\!]$.

By Lemma 12, the statement $w' \notin [\![\varphi]\!]$ in item 3 above implies that

$$w' \notin [\![\mathsf{S}_a\varphi]\!]. \tag{13}$$

On the other hand, because n is the $[\![\varphi]\!]$-achievement point such that $w \preceq n$, by Lemma 13, the assumption $w \in [\![\mathsf{S}_a\varphi]\!]$ implies that

$$w'' \in [\![\mathsf{S}_a\varphi]\!] \tag{14}$$

for each outcome w'' such that $w'' \preceq n$. Hence, by Definition 2, Eqs. 13, 14, and items 2 and 3 above imply that node n is a $[\![\mathsf{S}_a\varphi]\!]$-achievement point by agent a. Thus, by Lemma 1 and item 1 above, there is no $[\![\mathsf{S}_a\varphi]\!]$-achievement point m by agent b such that $w \preceq m$. Therefore, $w \notin [\![\mathsf{S}_b\mathsf{S}_a\varphi]\!]$ by item 5b of Definition 4. □

It is worth mentioning that the deliberative STIT modality shows the same trivialisation property: an agent never deliberatively sees to it that another agent deliberatively sees to something [71].

Finally, let us give an example of a non-trivial behaviour of the combination $\mathsf{C}_a\mathsf{C}_b$. In outcome w_2 of the game depicted in Fig. 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: $w_2 \in [\![\mathsf{C}_b\mathsf{C}_g$"Beach is left in prison"$]\!]$.

5.3 Mixed-Form Higher-Order Responsibility

Finally, let us consider the case of nesting two different forms of responsibility. In this subsection, we show that the combination $\mathsf{S}_a\mathsf{C}_b$ trivialises, while $\mathsf{C}_a\mathsf{S}_b$ does not.

First, let us show that $\mathsf{S}_a\mathsf{C}_b$ trivialises by proving that an agent is never responsible for seeing to a counterfactual responsibility of another agent. We prove this in Proposition 3 below. To increase the readability of its proof, we first show the next lemma.

Lemma 14 *For any formula $\varphi \in \Phi$, any node n, and any distinct agents $a, b \in \mathcal{A}$, if $n \in win_a([\![\varphi]\!])$, then $n \notin win_b([\![\neg\varphi]\!])$.*

Proof We prove this lemma by backward induction in the game tree, from the leaf nodes up to the root.

If node n is a leaf node, then the assumption $n \in win_a([\![\varphi]\!])$ of the lemma implies that $n \in [\![\varphi]\!]$ by item 1 and the minimality condition in Definition 3. Then, $n \notin [\![\neg\varphi]\!]$ by item 2 of Definition 4. Therefore, $n \notin win_b([\![\neg\varphi]\!])$ again by item 1 and the minimality condition in Definition 3.

If node n is a non-leaf node, then n is labelled by either agent a or some other agent. In the first case, by item 2 of Definition 3, the assumption $n \in win_a([\![\varphi]\!])$ of the lemma implies that $m \in win_a([\![\varphi]\!])$ for some child node m of node n. Then, $m \notin win_b([\![\neg\varphi]\!])$ by the induction hypothesis. Therefore, $n \notin win_b([\![\neg\varphi]\!])$ by item 3 of Definition 3.

In the second case, by item 3 of Definition 3, the assumption $n \in win_a([\![\varphi]\!])$ of the lemma implies that $l \in win_a([\![\varphi]\!])$ for each child node l of node n. Hence, $l \notin win_b([\![\neg\varphi]\!])$ by the induction hypothesis. Therefore, $n \notin win_b([\![\neg\varphi]\!])$ by items 2 and 3 of Definition 3. □

Proposition 3 *For any formula $\varphi \in \Phi$ and any distinct agents $a, b \in \mathcal{A}$, formula $\mathsf{S}_b\mathsf{C}_a\varphi$ is semantically equivalent to $\bot$.*

Proof By Definition 5, it suffices to prove $[\![\mathsf{S}_b\mathsf{C}_a\varphi]\!] = \varnothing$ for each extensive form game. Moreover, $[\![\mathsf{S}_b\mathsf{C}_a\varphi]\!] \subseteq [\![\mathsf{C}_a\varphi]\!]$ by Lemma 12. Thus, it suffices to show that $w \notin [\![\mathsf{S}_b\mathsf{C}_a\varphi]\!]$ for each outcome $w \in [\![\mathsf{C}_a\varphi]\!]$.

Consider an outcome $w \in [\![\mathsf{C}_a\varphi]\!]$. By item 4 of Definition 4, there exists a node n such that $w \preceq n$ and

$$n \in win_a([\![\neg\varphi]\!]). \tag{15}$$

Meanwhile, $[\![\varphi]\!] \supseteq [\![\mathsf{C}_a\varphi]\!]$ by item 4 of Definition 4. Then, $[\![\neg\varphi]\!] \subseteq [\![\neg\mathsf{C}_a\varphi]\!]$ by item 2 of Definition 4. Hence, by Lemma 11,

$$win_a([\![\neg\varphi]\!]) \subseteq win_a([\![\neg\mathsf{C}_a\varphi]\!]). \tag{16}$$

Then, $n \in win_a([\![\neg\mathsf{C}_a\varphi]\!])$ by Eqs. 15 and 16. Thus, $n \notin win_b([\![\mathsf{C}_a\varphi]\!])$ by Lemma 14 and item 2 of Definition 4. Together with the statement $w \preceq n$ and item 5a of Definition 4, it can be concluded that $w \notin [\![\mathsf{S}_b\mathsf{C}_a\varphi]\!]$. □

To see that the combination $\mathsf{C}_a\mathsf{S}_b$ does not trivialise, let us consider outcome w_2 of the game depicted in Fig. 1a. Recall from our earlier discussion that, in this outcome,

the Governor is responsible for seeing to it that Beach is left in prison. However, the Board could have *prevented this responsibility* by denying (D) the petition. Thus, in outcome w_2, the Board is counterfactually responsible for the Governor's responsibility for seeing to it that Beach is left in prison: $w_2 \in [\![\mathsf{C}_b\mathsf{S}_g\text{“Beach is left in prison”}]\!]$.

6 Responsibility Gap

In the past two decades, one of the important topics discussed in the ethics literature is the *responsibility gap* [8, 23–30]. The central question in this debate is, if something happens, is there always an agent that can be held responsible for it? If no, then a responsibility gap exists, which is undesirable in most situations. In this section, we study if the two forms of responsibility discussed in this paper are enough to avoid responsibility gaps in extensive form games. Note that, as discussed in Subsection 2.4, nobody should be responsible for a trivial truth. Hence, in this section, we only consider the responsibility gaps for statements that are *not* trivially true.

Let us go back to the example depicted in Fig. 1a. Recall that, in outcome w_1 where Beach is left in prison, the Board is responsible for seeing to it; in outcome w_2 where Beach is also left in prison, the Governor is responsible for seeing to it and also counterfactually responsible for it; in outcome w_3 where Beach is set free, 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.

In the rest of this section, we will study if responsibility gaps exist in arbitrary extensive form games. Let us start, however, by formally defining the ***gap formulae*** $\mathsf{G}^{\mathsf{c}}(\varphi)$, $\mathsf{G}^{\mathsf{s}}(\varphi)$, and $\mathsf{G}^{\mathsf{c},\mathsf{s}}(\varphi)$ for any formula $\varphi \in \Phi$. Informally, the counterfactual gap formula $\mathsf{G}^{\mathsf{c}}(\varphi)$ means that φ is true and no agent is counterfactually responsible for it, the seeing-to-it gap formula $\mathsf{G}^{\mathsf{s}}(\varphi)$ means that φ is true and no agent is responsible for seeing to it, and the combined gap formula $\mathsf{G}^{\mathsf{c},\mathsf{s}}(\varphi)$ means that φ is true and no agent is either counterfactually responsible or responsible for seeing to it:

$$\mathsf{G}^{\mathsf{c}}(\varphi) := \varphi \wedge \bigwedge_{a\in\mathcal{A}} \neg\mathsf{C}_a\varphi; \tag{17}$$

$$\mathsf{G}^{\mathsf{s}}(\varphi) := \varphi \wedge \bigwedge_{a\in\mathcal{A}} \neg\mathsf{S}_a\varphi; \tag{18}$$

$$\mathsf{G}^{\mathsf{c},\mathsf{s}}(\varphi) := \varphi \wedge \bigwedge_{a\in\mathcal{A}} \neg\mathsf{C}_a\varphi \wedge \bigwedge_{a\in\mathcal{A}} \neg\mathsf{S}_a\varphi. \tag{19}$$

6.1 Gaps in Games With Two Agents

In the extensive form games with only two agents, it is found that the two forms of responsibility discussed in this paper leave no responsibility gap. This result is formally stated in Theorem 3 below, whose proof uses the following lemma.

Lemma 15 *For any formula $\varphi \in \Phi$ and any node n in a two-agent extensive form game between agents a and b, if $n \notin win_a([\![\varphi]\!])$, then $n \in win_b([\![\neg\varphi]\!])$.*

Proof We prove this lemma by backward induction in the game tree, from the leaf nodes up to the root.

If node n is a leaf node, then the assumption $n \notin win_a([\![\varphi]\!])$ of the lemma implies that $n \notin [\![\varphi]\!]$ by item 1 of Definition 3. This further implies $n \in [\![\neg\varphi]\!]$ by item 2 of Definition 4. Therefore, $n \in win_b([\![\neg\varphi]\!])$ by item 1 of Definition 3.

If n is a non-leaf node, then n is labelled by either agent a or agent b. In the first case, by item 2 of Definition 3, the assumption $n \notin win_a([\![\varphi]\!])$ of the lemma implies that $m \notin win_a([\![\varphi]\!])$ for each child node m of node n. Then, $m \in win_b([\![\neg\varphi]\!])$ by the induction hypothesis. Therefore, $n \in win_b([\![\neg\varphi]\!])$ by item 3 of Definition 3.

In the second case, by item 3 of Definition 3, the assumption $n \notin win_a([\![\varphi]\!])$ of the lemma implies that there is a child node m of node n such that $m \notin win_a([\![\varphi]\!])$. Hence, $m \in win_b([\![\neg\varphi]\!])$ by the induction hypothesis. Therefore, $n \in win_b([\![\neg\varphi]\!])$ by item 2 of Definition 3. □

Together with Lemma 14, the above lemma shows that, at each node of an extensive form game between agents a and b, either agent a has a strategy to achieve a statement φ, or agent b has a strategy to achieve the negative statement $\neg\varphi$.

Theorem 3 *For any formula $\varphi \in \Phi$ and any two-agent extensive form game G, if $[\![\varphi]\!] \neq \Omega(G)$, then $[\![\mathsf{G}^{c,s}(\varphi)]\!] = \varnothing$.*

Proof Note that $\mathsf{G}^{c,s}(\varphi) = \mathsf{G}^{c}(\varphi) \wedge \mathsf{G}^{s}(\varphi)$ by Eqs. 17, 18, and 19. Then, $[\![\mathsf{G}^{c,s}(\varphi)]\!] = [\![\mathsf{G}^{c}(\varphi)]\!] \cap [\![\mathsf{G}^{s}(\varphi)]\!]$ by item 3 of Definition 4. Thus, to prove $[\![\mathsf{G}^{c,s}(\varphi)]\!] = \varnothing$, it is enough to show that $w \notin [\![\mathsf{G}^{c}(\varphi)]\!]$ for each outcome $w \in [\![\mathsf{G}^{s}(\varphi)]\!]$. Then, by Eq. 17 and items 2 and 3 of Definition 4, it suffices to show that for each outcome $w \in [\![\mathsf{G}^{s}(\varphi)]\!]$ there is an agent a such that $w \in [\![\mathsf{C}_a\varphi]\!]$.

By Eq. 18 and items 2 and 3 of Definition 4, the statement $w \in [\![\mathsf{G}^{s}(\varphi)]\!]$ implies that

$$w \in [\![\varphi]\!] \tag{20}$$

and

$$w \notin [\![\mathsf{S}_b\varphi]\!] \tag{21}$$

for each agent b in game G. Meanwhile, by Eq. 20, Lemma 1, and the assumption $[\![\varphi]\!] \neq \Omega(G)$ of the lemma, there exists an agent b and a $[\![\varphi]\!]$-achievement point n by agent b such that $w \preceq n$. This means item 5b of Definition 4 is true for the agent b toward outcome w. Thus, to make Eq. 21 true, item 5a of Definition 4 must be false. Hence, there is a node m such that $w \preceq m$ and $m \notin win_b([\![\varphi]\!])$. Since G is a two-agent game, let a be the agent in the game distinct from agent b. Then, $m \in win_a([\![\neg\varphi]\!])$ by Lemma 15. Therefore, $w \in [\![\mathsf{C}_a\varphi]\!]$ by item 4 of Definition 4, Eq. 20, and the fact that $w \preceq m$. □

Theorem 3 shows that the undesirable responsibility gap should not be a concern if there are only two agents in the system. The result further applies to more general settings where more agents might exist but only two of them are relevant to the decision.

6.2 Gaps in Games With More Than Two Agents

To see if the responsibility gaps exist in extensive form games with *more than* two agents, let us go back to the story of Beach's clemency, which is not as simple as we 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 [72].

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 [73]. Figure 5 depicts the extensive form game that captures the situation after the bill was introduced. If the Montana State Legislature (agent l) rejects (H) the bill, then the game continues as in Fig. 1a. If the Legislature approves (I) the bill, then the Governor unilaterally decides whether to grant the clemency.

By Definition 2, Definition 3, and Definition 4, it is easily observable that, 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, who sees to Beach being left in prison in the achievement way in outcome w_1, 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 Eq. 19,

$$[\![\mathsf{G}^{c,s}(\text{“Beach is left in prison”})]\!] = \{w_1\}.$$

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.

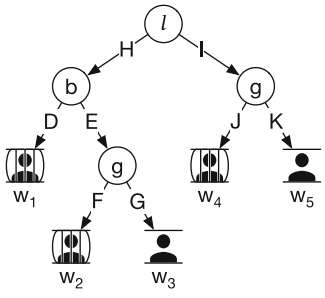


Fig. 5 Barry Beach's case of clemency

6.3 Hierarchy of Responsibility Gaps

As seen in the previous subsection, the responsibility gaps may exist in an extensive form game. A further question is whether 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 Fig. 5, it is the Legislature that is counterfactually responsible for the gap in outcome w_1. Indeed, the Legislature could prevent the gap formula $\mathsf{G}^{\mathsf{c},\mathsf{s}}$("Beach is left in prison") from being true by approving (I) the bill:

$$w_1 \in [\![\mathsf{C}_l \mathsf{G}^{\mathsf{c},\mathsf{s}}(\text{"Beach is left in prison"})]\!].$$

In addition, in this example, the Board is also counterfactually responsible for the gap in outcome w_1. That is, $w_1 \in [\![\mathsf{C}_b \mathsf{G}^{\mathsf{c},\mathsf{s}}(\text{"Beach is left in prison"})]\!]$.

Next, let us consider another condition where no agent is responsible for a gap. By *second-order gap* for a formula φ we mean the presence of outcomes in which $\mathsf{G}^{\mathsf{c},\mathsf{s}}(\varphi)$ is true but no agent 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 $i \geq 0$, define the i^{th}-order (combined) gap formula $\mathsf{G}_i^{\mathsf{c},\mathsf{s}}(\varphi)$ recursively as:

$$\mathsf{G}_i^{\mathsf{c},\mathsf{s}}(\varphi) := \begin{cases} \mathsf{G}_{i-1}^{\mathsf{c},\mathsf{s}}(\varphi) \wedge \bigwedge_{a\in\mathcal{A}} \neg \mathsf{C}_a \mathsf{G}_{i-1}^{\mathsf{c},\mathsf{s}}(\varphi) \\ \qquad\qquad \wedge \bigwedge_{a\in\mathcal{A}} \neg \mathsf{S}_a \mathsf{G}_{i-1}^{\mathsf{c},\mathsf{s}}(\varphi), & i \geq 1; \\ \varphi, & i = 0. \end{cases} \tag{22}$$

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

$$\mathsf{G}_i^{\mathsf{c}}(\varphi) := \begin{cases} \mathsf{G}_{i-1}^{\mathsf{c}}(\varphi) \wedge \bigwedge_{a\in\mathcal{A}} \neg \mathsf{C}_a \mathsf{G}_{i-1}^{\mathsf{c}}(\varphi), & i \geq 1; \\ \varphi, & i = 0. \end{cases} \tag{23}$$

One can similarly define the i^{th}-order seeing-to-it gap formula $\mathsf{G}_i^{\mathsf{s}}(\varphi)$. It is easy to see from Eqs. 17, 18, and 19 that the first order gap statements $\mathsf{G}_1^{\mathsf{c},\mathsf{s}}(\varphi)$, $\mathsf{G}_1^{\mathsf{c}}(\varphi)$, and $\mathsf{G}_1^{\mathsf{s}}(\varphi)$ are equivalent to the previously discussed gap formulae $\mathsf{G}^{\mathsf{c},\mathsf{s}}(\varphi)$, $\mathsf{G}^{\mathsf{c}}(\varphi)$, and $\mathsf{G}^{\mathsf{s}}(\varphi)$, respectively.

Perhaps the higher-order gaps are less intuitive, but they do exist. As an example, for an arbitrary integer $i > 1$, consider a situation when a manager a of a plant gets an important order for a product that could be assembled at either of the two assembly lines at the plant. The assembly takes $i+1$ steps. The first assembly line is manned by workers $b_1, \ldots, b_{i+1}$ and the second by workers $c_1, \ldots, c_{i+1}$. None of the workers is very reliable and each of them can make a mistake that cannot be fixed. Figure 6 represents this scenario as an extensive form game between agents $a, b_1, \ldots, b_{i+1}, c_1, \ldots, c_{i+1}$.

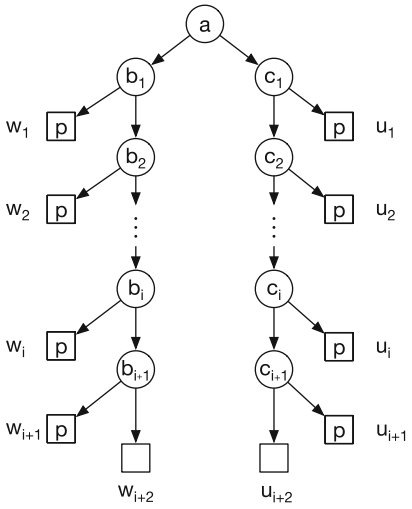


Fig. 6 A game in which the set $[\![\mathsf{G}_i^{\mathsf{c},\mathsf{s}}(p)]\!]$ is not empty

Propositional variable p represents the statement "product is damaged beyond repair". This game consists of

- $2i + 3$ non-leaf nodes, each labelled with a distinct agent;
- $2i + 4$ outcomes: $w_1, \ldots, w_{i+2}, u_1, \ldots, u_{i+2}$, among which, w_{i+2} and u_{i+2} are labelled with the empty set while the others are labelled with the set $\{p\}$.

We will show that the i^{th}-order gap exists in this game by showing that the truth set $[\![\mathsf{G}_i^{\mathsf{c},\mathsf{s}}(p)]\!]$ is not empty.

Note that, by Eq. 22,

$$[\![\mathsf{G}_0^{\mathsf{c},\mathsf{s}}(p)]\!] = [\![p]\!] = \{w_1, \ldots, w_{i+1}, u_1, \ldots, u_{i+1}\}.$$

It is easily observable that no agent in this game has an upfront strategy to guarantee that the game ends with an outcome where p is true. That is, the root node does *not* belong to the set $win_g([\![p]\!])$ for any agent g in this game. Thus, by item 5a of Definition 4, no agent is responsible for seeing to p in any of the outcomes. At the same time, for all the outcomes where p is true, only in w_{i+1} and u_{i+1} statement p can be prevented on the path of play (by agents b_{i+1} and c_{i+1}, respectively). Hence, by item 4 of Definition 4, no agent is counterfactually responsible for p in outcomes $w_1, \ldots, w_i, u_1, \ldots, u_i$. As a result,

$$[\![\mathsf{G}_1^{\mathsf{c},\mathsf{s}}(p)]\!] = \{w_1, \ldots, w_i, u_1, \ldots, u_i\}$$

by Eq. 22. With the same reasoning process, it is not hard to deduce the following statements:

$$\begin{aligned}
&[\![\mathsf{G}_2^{c,s}(p)]\!] = \{w_1, \ldots, w_{i-1}, u_1, \ldots, u_{i-1}\};\\
&[\![\mathsf{G}_3^{c,s}(p)]\!] = \{w_1, \ldots, w_{i-2}, u_1, \ldots, u_{i-2}\};\\
&\quad\vdots\\
&[\![\mathsf{G}_k^{c,s}(p)]\!] = \{w_1, \ldots, w_{i+1-k}, u_1, \ldots, u_{i+1-k}\};\\
&\quad\vdots\\
&[\![\mathsf{G}_i^{c,s}(p)]\!] = \{w_1, u_1\}.
\end{aligned}$$

The last formula above shows that, in outcomes w_1 and u_1 of the game depicted in Fig. 6, the i^{th}-order responsibility gap exists[11]. Recall that i is an arbitrary integer greater than 1. In conclusion, no matter how high the order we consider, there always exists an extensive form game where the higher-order gap exists.

Despite this, it can be proved that, *for any given extensive form game, the higher-order responsibility gap does not exist if a sufficiently high order is considered*. This is formally captured in Theorem 4 and Corollary 1 below, which claim that, for any extensive form game and any formula $\varphi \in \Phi$ that is not a trivial truth, the sets $[\![\mathsf{G}_i^c(\varphi)]\!]$ and $[\![\mathsf{G}_i^{c,s}(\varphi)]\!]$ are empty for large enough integer i. To increase the readability of the proofs, let us first prove two lemmas. These lemmas show that the set $[\![\mathsf{G}_i^c(\varphi)]\!]$ monotonously shrinks to empty as the order i increases.

Lemma 16 $[\![\mathsf{G}_{i+1}^c(\varphi)]\!] \subseteq [\![\mathsf{G}_i^c(\varphi)]\!]$ *for any formula* $\varphi \in \Phi$ *and any integer* $i \geq 0$.

Proof The statement of the lemma follows from Eq. 23 and item 3 of Definition 4. □

Lemma 17 *For any formula* $\varphi \in \Phi$, *any integer* $i \geq 0$, *and any extensive form game* G, *if* $\varnothing \subsetneq [\![\mathsf{G}_i^c(\varphi)]\!] \subsetneq \Omega(G)$, *then* $[\![\mathsf{G}_{i+1}^c(\varphi)]\!] \subsetneq [\![\mathsf{G}_i^c(\varphi)]\!]$.

Proof The assumption $\varnothing \subsetneq [\![\mathsf{G}_i^c(\varphi)]\!] \subsetneq \Omega(G)$ of the lemma, by item 2 of Definition 4, implies that $\varnothing \subsetneq [\![\neg\mathsf{G}_i^c(\varphi)]\!] \subsetneq \Omega(G)$. Then, on the one hand, there is an outcome $w \in [\![\neg\mathsf{G}_i^c(\varphi)]\!]$. On the other hand, by Lemma 1, there is an $[\![\neg\mathsf{G}_i^c(\varphi)]\!]$-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 $w' \notin [\![\neg\mathsf{G}_i^c(\varphi)]\!]$;
3. $w'' \in [\![\neg\mathsf{G}_i^c(\varphi)]\!]$ for each outcome w'' such that $w'' \preceq n$.

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

$$parent(n) \in win_a([\![\neg\mathsf{G}_i^c(\varphi)]\!]). \tag{24}$$

[11] If we consider another statement "agent b_1 damaged the product", which is true only in outcome w_1, then agent b_1 is counterfactually responsible for it in outcome w_1. In other words, in outcome w_1, although the i^{th}-order gap for the *damage of the product* exists, agent b_1 is counterfactually responsible for *herself damaging the product*.

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

$$w' \in [\![\mathsf{G}_i^{\mathsf{c}}(\varphi)]\!]. \tag{25}$$

Thus, $w' \in [\![\mathsf{C}_a \mathsf{G}_i^{\mathsf{c}}(\varphi)]\!]$ by the part $w' \preceq parent(n)$ of item 2 above, Eq. 24, and item 4 of Definition 4. Then, $w' \notin [\![\neg \mathsf{C}_a \mathsf{G}_i^{\mathsf{c}}(\varphi)]\!]$ by item 2 of Definition 4. This further implies that $w' \notin [\![\mathsf{G}_{i+1}^{\mathsf{c}}(\varphi)]\!]$ by Eq. 23 and item 3 of Definition 4. Hence, $[\![\mathsf{G}_{i+1}^{\mathsf{c}}(\varphi)]\!] \neq [\![\mathsf{G}_i^{\mathsf{c}}(\varphi)]\!]$ by Eq. 25. Therefore, $[\![\mathsf{G}_{i+1}^{\mathsf{c}}(\varphi)]\!] \subsetneq [\![\mathsf{G}_i^{\mathsf{c}}(\varphi)]\!]$ by Lemma 16. □

Theorem 4 *$[\![\mathsf{G}_i^{\mathsf{c}}(\varphi)]\!] = \varnothing$ for each extensive form game G, each integer $i \geq |\Omega(G)| - 1$, and each formula $\varphi \in \Phi$ such that $[\![\varphi]\!] \subsetneq \Omega(G)$.*

Proof Suppose the opposite, then there is an integer $j \geq |\Omega(G)| - 1$ such that

$$[\![\mathsf{G}_j^{\mathsf{c}}(\varphi)]\!] \neq \varnothing. \tag{26}$$

Note that, by Eq. 23, Lemma 16, and the assumption $[\![\varphi]\!] \subsetneq \Omega(G)$,

$$\Omega(G) \supsetneq [\![\varphi]\!] = [\![\mathsf{G}_0^{\mathsf{c}}(\varphi)]\!] \supseteq [\![\mathsf{G}_1^{\mathsf{c}}(\varphi)]\!] \supseteq [\![\mathsf{G}_2^{\mathsf{c}}(\varphi)]\!] \supseteq \dots \tag{27}$$

Then, $\varnothing \subsetneq [\![\mathsf{G}_i^{\mathsf{c}}(\varphi)]\!] \subsetneq \Omega(G)$ for each integer i such that $0 \leq i \leq j$ by Eq. 26. Thus, $|[\![\mathsf{G}_i^{\mathsf{c}}(\varphi)]\!]| - |[\![\mathsf{G}_{i+1}^{\mathsf{c}}(\varphi)]\!]| \geq 1$ for each integer $i \leq j$ by Lemma 17. Hence, $|[\![\mathsf{G}_0^{\mathsf{c}}(\varphi)]\!]| - |[\![\mathsf{G}_{j+1}^{\mathsf{c}}(\varphi)]\!]| \geq j + 1$. Note that, $|[\![\mathsf{G}_{j+1}^{\mathsf{c}}(\varphi)]\!]| \geq 0$ and $j \geq |\Omega(G)| - 1$. Thus, $|[\![\mathsf{G}_0^{\mathsf{c}}(\varphi)]\!]| \geq |\Omega(G)|$, which contradicts Eq. 27. □

Given the above theorem, to prove Corollary 1 below, it suffices to show the statement in the next lemma is true.

Lemma 18 *$[\![\mathsf{G}_i^{\mathsf{c,s}}(\varphi)]\!] \subseteq [\![\mathsf{G}_i^{\mathsf{c}}(\varphi)]\!]$ for any formula $\varphi \in \Phi$ and any integer $i \geq 0$.*

Proof We prove the lemma by induction on integer i. If $i = 0$, then $\mathsf{G}_i^{\mathsf{c,s}}(\varphi) = \varphi = \mathsf{G}_i^{\mathsf{c}}(\varphi)$ by Eqs. 22 and 23. Therefore, $[\![\mathsf{G}_i^{\mathsf{c,s}}(\varphi)]\!] = [\![\mathsf{G}_i^{\mathsf{c}}(\varphi)]\!]$.

In the cases where $i \geq 1$, by Eqs. 22, 23 and item 3 of Definition 4,

$$[\![\mathsf{G}_i^{\mathsf{c,s}}(\varphi)]\!] \subseteq [\![\mathsf{G}_{i-1}^{\mathsf{c,s}}(\varphi)]\!] \cap \bigcap_{a \in \mathcal{A}} [\![\neg \mathsf{C}_a \mathsf{G}_{i-1}^{\mathsf{c,s}}(\varphi)]\!] \tag{28}$$

and

$$[\![\mathsf{G}_i^{\mathsf{c}}(\varphi)]\!] = [\![\mathsf{G}_{i-1}^{\mathsf{c}}(\varphi)]\!] \cap \bigcap_{a \in \mathcal{A}} [\![\neg \mathsf{C}_a \mathsf{G}_{i-1}^{\mathsf{c}}(\varphi)]\!]. \tag{29}$$

Consider an arbitrary outcome $w \in [\![\mathsf{G}_i^{\mathsf{c,s}}(\varphi)]\!]$. It suffices to show that $w \in [\![\mathsf{G}_i^{\mathsf{c}}(\varphi)]\!]$. Note that, by Eq. 28, the assumption $w \in [\![\mathsf{G}_i^{\mathsf{c,s}}(\varphi)]\!]$ implies that

$$w \in [\![\mathsf{G}_{i-1}^{\mathsf{c,s}}(\varphi)]\!] \tag{30}$$

and

$$w \in \bigcap_{a \in \mathcal{A}} [\![\neg \mathsf{C}_a \mathsf{G}^{\mathsf{c},\mathsf{s}}_{i-1}(\varphi)]\!]. \tag{31}$$

By the induction hypothesis, Eq. 30 further implies that $w \in [\![\mathsf{G}^{\mathsf{c}}_{i-1}(\varphi)]\!]$. Thus, by Eq. 29, to show that $w \in [\![\mathsf{G}^{\mathsf{c}}_{i}(\varphi)]\!]$, it suffices to prove that $w \in [\![\neg \mathsf{C}_a \mathsf{G}^{\mathsf{c}}_{i-1}(\varphi)]\!]$ for each agent $a \in \mathcal{A}$.

Towards a contradiction, suppose $w \notin [\![\neg \mathsf{C}_a \mathsf{G}^{\mathsf{c}}_{i-1}(\varphi)]\!]$ for some agent a. Then, $w \in [\![\mathsf{C}_a \mathsf{G}^{\mathsf{c}}_{i-1}(\varphi)]\!]$ by item 2 of Definition 4. Hence, by item 4 of Definition 4, there exists a node n such that

$$w \preceq n \tag{32}$$

and

$$n \in win_a([\![\neg \mathsf{G}^{\mathsf{c}}_{i-1}(\varphi)]\!]). \tag{33}$$

At the same time, $[\![\neg \mathsf{G}^{\mathsf{c}}_{i-1}(\varphi)]\!] \subseteq [\![\neg \mathsf{G}^{\mathsf{c},\mathsf{s}}_{i-1}(\varphi)]\!]$ by the induction hypothesis and item 2 of Definition 4. Thus, by Lemma 11 and Eq. 33,

$$n \in win_a([\![\neg \mathsf{G}^{\mathsf{c},\mathsf{s}}_{i-1}(\varphi)]\!]).$$

Hence, $w \in [\![\mathsf{C}_a \mathsf{G}^{\mathsf{c},\mathsf{s}}_{i-1}(\varphi)]\!]$ by Eqs. 30, 32, and item 4 of Definition 4, which contradicts Eq. 31. □

The next corollary follows from the above lemma and Theorem 4.

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

Although a gap of an arbitrarily high order would technically exist as the example in Fig. 6 illustrates, Theorem 4 and Corollary 1 show that, for a given system (*i.e.* extensive form game), there always is a high enough order such that the gap of this order is empty. Particularly, this gap-free order is less than the number of potential outcomes of this system.

It is worth mentioning that, in the literature, a widely-discussed solution to the responsibility gap problem is to consider collective responsibility for a group of agents [69, 74–76]. That is, in a situation when responsibility could not be attributed to any single agent, a coalition of several agents might be responsible. However, this approach sometimes causes concern due to the lack of reason why several agents should be treated together as a *single* subject of responsibility [77–80] (*e.g.* when common knowledge, intention, goal, or communication is missing). Our higher-order gap-free observation brings an alternative approach to the responsibility gap problem in settings where more than two agents are involved in the decision.

7 Challenges in Imperfect Information Settings

In the above discussion, we consider only the multi-step decision schemes that can be modelled as extensive form games from Definition 1. In that definition, informally,

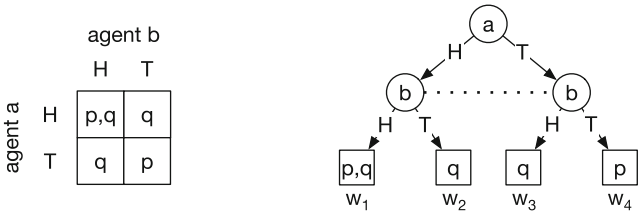
we assumed that the agents always know the current state (*i.e.* the real path of play decided by all actions having been taken so far) whenever they need to make a choice. This is usually called ***perfect information*** setting. However, this assumption does not always hold. For example, in the matching pennies game between agents a and b (see the game matrix in Fig. 7a), the two agents concurrently choose either *Head* (H) or *Tail* (T). In this setting, none of the agents knows the choice of the other one before making her own choice.

Let us suppose agent a makes her decision one picosecond before agent b does so. In a sense, the two agents still decide concurrently because nothing is expected to happen in such a short interval. Then, the matching pennies game can also be denoted by a game tree shown in Fig. 7b. In this tree, agent a first decides *Head* or *Tail*, then agent b decides *Head* or *Tail* without knowing agent a's choice. Informally speaking, when agent b makes the decision, she only knows that the current state is one of these two b-labelled nodes but does not know which of them is the right one. In this situation, it is said that agent b cannot *distinguish* these two b-labelled nodes. In the literature, the indistinguishability is usually denoted by a dashed line between the indistinguishable nodes in the game tree (*e.g.* the dashed line between the two b-labelled nodes in Fig. 7b). In general, an extensive form game is called ***imperfect information*** if there are some indistinguishable nodes for at least one agent in the game tree.

7.1 Failure of the Existing Definitions

To see how the two notions of responsibility defined in items 4 and 5 of Definition 4 work in imperfect information settings, let us consider outcome w_1 in Fig. 7b, where the choices of the two agents match each other. Let propositional variable p represent "match" and be true in outcomes w_1 and w_4. If the indistinguishability of the two b-labelled nodes is ignored, then, by Definition 2, outcome w_1 is the $[\![p]\!]$-achievement point by agent b; by Definition 3, all the three non-leaf nodes are in both the set $win_b([\![p]\!])$ and the set $win_b([\![\neg p]\!])$. Then, by items 4 and 5 of Definition 4, agent b should be both counterfactually responsible and responsible for seeing to p in outcome w_1.

However, it is easily observable that agent b cannot prevent p on the path of play to outcome w_1. Intuitively, this is because agent b does not know the choice of agent a



(a) game matrix (b) game tree

Fig. 7 Matching pennies game

and thus cannot guarantee match or mismatch. Formally, even though both b-labelled nodes in Fig. 7b are in the set $win_b([\![\neg p]\!])$, the strategies to achieve $\neg p$ are different in these two nodes. However, in order to guarantee $\neg p$ in the outcome, agent b needs a ***uniform strategy*** (*i.e.* to act in the same way) in the two b-labelled nodes because she cannot distinguish them. In a word, *agent b cannot prevent p on the path of play to outcome* w_1 due to the lack of knowledge of agent a's choice. For this reason, it is improper to hold agent b counterfactually responsible for p in outcome w_1.

Similarly, agent b has no uniform strategy to achieve p in each of the three non-leaf nodes in Fig. 7b. In this sense, *agent b does not have an upfront ability to achieve p*. Also, when agent b chooses *Head* in the left b-labelled node, she does not know that p would be unavoidable, again because she cannot distinguish the two b-labelled nodes and choosing *Head* in the right b-labelled node will result in outcome w_3 and let p be false. In other words, on the path of play to outcome w_1, *when agent b eliminates the last possibility for* $\neg p$*, she does not know that*. As a result, it is also improper to hold agent b responsible for seeing to p in outcome w_1.

As shown in the above example, the original definitions in items 4 and 5 of Definition 4 fail to properly capture the counterfactual responsibility and the seeing-to-it responsibility in imperfect information settings. It is also worth noting that, in the matching pennies game as shown in Fig. 7b, a proper definition should ascribe neither agent a nor agent b counterfactually responsible for p or responsible for seeing to p. In other words, a responsibility gap for p exists in outcomes w_1 and w_4 of this game, which means $[\![\mathsf{G}^{\mathsf{c},\mathsf{s}}(p)]\!] = \{w_1, w_4\}$. Then, $[\![\mathsf{G}^{\mathsf{c},\mathsf{s}}(p)]\!] = [\![p]\!]$ by the fact that $[\![p]\!] = \{w_1, w_4\}$. Thus, by Eq. 22,

$$[\![\mathsf{G}_i^{\mathsf{c},\mathsf{s}}(p)]\!] = \{w_1, w_4\}$$

for each integer $i \geq 0$. This example shows that, *in imperfect information settings, the higher-order responsibility gap may always exist no matter how high the order is, even in games with only two agents*. This result contrasts with Theorem 3, which says the gap does not exist in two-agent extensive form games with perfect information. It also contrasts with Corollary 1, which says a higher-order responsibility gap can always be filled in perfect information settings if enough high order is considered.

7.2 Extension of the Original Notions

In imperfect information settings, the *strategic ability* of an agent to achieve a statement φ is usually captured by a uniform strategy of the agent that guarantees φ in the outcome. This is exactly how counterfactual responsibility is defined in imperfect information settings in [40, 41]. More specifically, to hold an agent counterfactually responsible in imperfect information settings, a uniform strategy of the agent to achieve a negative condition should exist.

In the same way, let us consider the notion of seeing-to-it responsibility. As discussed in Subsection 2.4.3, this notion is the combination of backwards-strategic seeing-to-it and achievement seeing-to-it. The extension of the backwards-strategic seeing-to-it into imperfect information settings seems to be straightforward in the sense

that both the upfront ability and its maintenance can be interpreted as the existence of a uniform strategy. On the contrary, the situation with the achievement seeing-to-it is more complicated. Note that the notion of achievement seeing-to-it is taken to denote the "responsible action" of an agent that guarantees a statement. According to its original idea of "eliminating the last possibility for a negative condition", this notion should remain the same in imperfect information settings as in perfect information settings. In this sense, on the path of play to outcome w_1 of the game in Fig. 7b, although both agents choose *Head* concurrently, the action of agent b is a responsible action while the same action of agent a is not.

The situation will be even more confusing if the seeing-to-it responsibility is defined as the combination of the extended notion of backwards-strategic seeing-to-it and the original notion of achievement seeing-to-it. For instance, consider another proposition q, which represents "*Head* appears" in the game in Fig. 7b. It can be observed that, in outcome w_1, both agents see to statement q backwards-strategically, while only agent a sees to q in the achievement way. As a consequence, agent a is responsible for seeing to q but agent b is not after they choose *Head* concurrently. However, looking back at the matrix in Fig. 7a, by choosing *Head*, both agents should be responsible for seeing to q, which is in line with the discussion of the seeing-to-it responsibility in strategic games [12].

To hold agent b also responsible for seeing to q in outcome w_1 of the tree in Fig. 7b, a possible solution is to extend the notion of achievement seeing-to-it to ***knowingly-achievement seeing-to-it*** in imperfect information settings. Informally, an agent a is said to see to statement φ in the knowingly-achievement way if *(1) she takes an action H at a node n on the path of play, (2) from at least one indistinguishable node of n, a negative condition is still available, and, (3) in each of the indistinguishable nodes of node n, action H eliminates the last possibility for $\neg\varphi$ if existing*. In short, the notion of knowingly-achievement seeing-to-it means that an agent takes an action, making φ unavoidable henceforth from all the indistinguishable states. With the extended notion, in Fig. 7b, the left b-labelled node is the knowingly-$[\![q]\!]$-achievement point by agent a and outcome w_1 is the knowingly-$[\![q]\!]$-achievement point by agent b. Then, both agents can be held responsible for seeing to q in outcome w_1.

Although the knowingly-achievement seeing-to-it solves the issue in the matching pennies game, it also arouses other concerns. Let us consider another situation starting with an insect lying motionlessly on the floor. As shown in Fig. 8, agent a, the *Nature*, decides if the insect is *dead* (D) or *sleeping* (S), and then agent b, a human who cannot distinguish these two states, finds the insect and decides whether to *apply* (A) insecticide or to *leave* (L) the insect alone. Propositional variable q represents that the

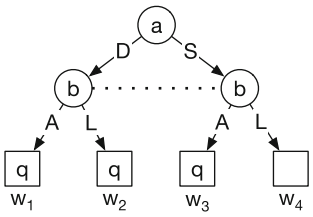


Fig. 8 A game setting where the knowingly-achievement seeing-to-it fails

insect is dead and is true in outcomes w_1, w_2, and w_3. Consider outcome w_1 where the insect is dead by the decision of agent a and then, without knowing this, agent b applies insecticide. It can be observed that, in outcome w_1 of the game in Fig. 8, agent b sees to the death of the insect both backwards-strategically and in the knowingly-achievement way and thus should be responsible for seeing to the death of the insect. Namely, agent b is said to be responsible for the death of an insect that is *already* dead when she found it, which seems counterintuitive. Generally speaking, when taken as the action part of the responsibility, the notion of knowingly-achievement seeing-to-it may hold an action responsible when it does not have any effect on the real world.

One may further add the requirement of an "actual elimination" into the notion of knowingly-achievement seeing-to-it. Let us call it ***consciously-achievement seeing-to-it***. Informally, an agent is said to see to statement φ in the consciously-achievement way if *(1) she takes an action H at a node n on the path of play (2) from node n, a negative condition is still available, and, (3) in each of the indistinguishable nodes of node n, action H eliminates the last possibility for $\neg\varphi$ if existing*. In other words, the consciously-achievement seeing-to-it means that an agent takes an action that makes φ unavoidable henceforth and the agent knows it when taking this action. However, it is easy to see that this notion fails again to hold agent b of the game in Fig. 7b responsible for seeing to statement q in outcome w_1.

In essence, we would like to claim that, *there is no proper definition of the seeing-to-it responsibility based on the tree structure in imperfect information extensive form games*. This is true because, as may have been noticed already, the game trees in Figs. 7b and 8 represent exactly the same game except for using different notations to represent the actions. However, agent b in this game tree should be responsible for seeing to statement q if the game tree represents the matching pennies game but should not be responsible if the game tree represents the dead-insect game. The following facts might be the cause of the above conflict:

1. Currently, the seeing-to-it responsibility is defined based on the tree structure of the extensive form games.
2. In imperfect information extensive form games, the information about the order of actions and the epistemic states of the agents is mixed in the tree structure and cannot be distinguished from each other.
3. The order of actions and the epistemic states of the agents affect the attribution of the seeing-to-it responsibility differently.

In a nutshell, to find a proper definition of the seeing-to-it responsibility in imperfect information settings, it is not enough to simply modify the notions. A modification of the tree structure may also be needed. Among other options, such a modification might include allowing events (such as the death of an insect) to happen in the middle of the game, not just in a leaf node.

8 Conclusion

The existing definitions of seeing-to-it modalities have clear shortcomings when viewed as possible forms of responsibility. In this paper, we made some revisions and

combined them into a single definition of seeing-to-it responsibility that addresses the shortcomings. By proving the undefinability results, we have shown that the proposed notion is semantically independent of the counterfactual responsibility already discussed in the literature. The discussion of higher-order responsibility that is easily expressed with our modal language offers a new perspective on the issue of the improper subject of responsibility. The other important contribution of this paper is the hierarchy of responsibility gaps. We believe that taking into account higher-order responsibilities is an important step towards responsibility attribution in complex multiagent settings such as hybrid human-machine systems. In particular, considering higher-order responsibility gaps is a key to the design of responsible agents and systems.

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 changed the clemency procedure. By doing so, they, perhaps unintentionally, prevented the potential responsibility gap existing in outcome w_1 of Fig. 5. In November of the same year, the Governor granted Beach a clemency [72].

Author Contributions The authors contributed equally to this paper.

Funding Qi Shi is supported by the China Scholarship Council (CSC No.202206070014).

Declarations

Conflicts of Interest The authors wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

References

1. Shi, Q. (2024). Responsibility in extensive form games. In: *Proceedings of the 38th AAAI Conference on Artificial Intelligence (pp. 19920–19928) (AAAI-24)*.
2. van de Poel, I. (2011). The relation between forward-looking and backward-looking responsibility. In: *Moral Responsibility: Beyond Free Will and Determinism,* (pp. 37–52). Springer, Dordrecht.
3. Goodin, R. E. (1995). *Utilitarianism as a Public Philosophy*. Cambridge: Cambridge University Press.
4. Yazdanpanah, V., Gerding, E. H., Stein, S., Dastani, M., Jonker, C. M., Norman, T. J., & Ramchurn, S. D. (2023). Reasoning about responsibility in autonomous systems: challenges and opportunities. *AI & Society, 38*(4), 1453–1464.
5. Hart, H. (2008 [1968]). *Punishment and Responsibility: Essays in the Philosophy of Law*. Oxford: Oxford University Press.
6. Halpern, J. Y. (2016). *Actual Causality*. Massachusetts: MIT Press.
7. Dastani, M., & Yazdanpanah, V. (2023). Responsibility of AI systems. *AI & Society, 38*(2), 843–852.

8. Matthias, A. (2004). The responsibility gap: ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*, 175–183.
9. Halpern, J. Y., & Kleiman-Weiner, M. (2018). Towards formal definitions of blameworthiness, intention, and moral responsibility. In: *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (pp. 1853–1860) (AAAI-18)*.
10. Bulling, N., & Dastani, M. (2013). Coalitional responsibility in strategic settings. In: *Proceedings of the 14th Workshop on Computational Logic in Multi-Agent Systems*, (pp. 172–189). Springer.
11. Yazdanpanah, V., & Dastani, M. (2016). Quantified degrees of group responsibility. In: *Proceedings of the 11th International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems (COIN-15)*, (pp. 418–436). Springer.
12. Naumov, P., & Tao, J. (2023). Counterfactual and seeing-to-it responsibilities in strategic games. *Annals of Pure and Applied Logic, 174*(10), 103353.
13. Sullins, J. P. (2006). When is a robot a moral agent? *International Review of Information Ethics, 6*, 23–30.
14. Stahl, B. C. (2006). Responsible computers? A case for ascribing quasi-responsibility to computers independent of personhood or agency. *Ethics and Information Technology, 8*, 205–213.
15. Pasupathi, M., & Wainryb, C. (2010). Developing moral agency through narrative. *Human Development, 53*(2), 55–80.
16. Parthemore, J., & Whitby, B. (2013). What makes any agent a moral agent? Reflections on machine consciousness and moral agency. *International Journal of Machine Consciousness, 5*(02), 105–129.
17. Parthemore, J., & Whitby, B. (2014). Moral agency, moral responsibility, and artifacts: What existing artifacts fail to achieve (and why), and why they, nevertheless, can (and do!) make moral claims upon us. *International Journal of Machine Consciousness, 6*(02), 141–161.
18. Cervantes, J.-A., López, S., Rodríguez, L.-F., Cervantes, S., Cervantes, F., & Ramos, F. (2020). Artificial moral agents: A survey of the current status. *Science and Engineering Ethics, 26*, 501–532.
19. Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics, 26*(4), 2051–2068.
20. Sebastián, M. Á. (2021). First-person representations and responsible agency in AI. *Synthese, 199*(3–4), 7061–7079.
21. Frankfurt, H. G. (1969). Alternate possibilities and moral responsibility. *The Journal of Philosophy, 66*(23), 829–839.
22. Broersen, J. (2011). Deontic epistemic STIT logic distinguishing modes of mens rea. *Journal of Applied Logic, 9*(2), 137–152.
23. Braham, M., & VanHees, M. (2011). Responsibility voids. *The Philosophical Quarterly, 61*(242), 6–15.
24. Duijf, H. (2018). Responsibility voids and cooperation. *Philosophy of the Social Sciences, 48*(4), 434–460.
25. Burton, S., Habli, I., Lawton, T., McDermid, J., Morgan, P., & Porter, Z. (2020). Mind the gaps: assuring the safety of autonomous systems from an engineering, ethical, and legal perspective. *Artificial Intelligence, 279*, 103201.
26. Gunkel, D. J. (2020). Mind the gap: responsible robotics and the problem of responsibility. *Ethics and Information Technology, 22*(4), 307–320.
27. Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & 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.
28. 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*, (pp. 390–400).
29. Braham, M., & Hees, M. (2018). Voids or fragmentation: moral responsibility for collective outcomes. *The Economic Journal, 128*(612), 95–113.
30. Hiller, S., Israel, J., & 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)*, (pp. 435–457). Springer.
31. US Supreme Court (1993). Herrera v. Collins, 506 U.S. 390. https://supreme.justia.com/cases/federal/us/506/390. Accessed 23 Dec 2023

32. 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
33. Constitution Convention (1889). Constitution of the State of Montana. https://courts.mt.gov/external/library/docs/1889cons.pdf. Accessed 14 May 2023
34. Montana Board of Pardons and Parole (2023). History. https://bopp.mt.gov/History. Accessed 14 May 2023
35. Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.
36. Harsanyi, J. C. (1967). Games with incomplete information played by "Bayesian" players, i-iii. part i. the basic model. *Management Science, 14*(3), 159–182
37. Osborne, M. J., & Rubinstein, A. (1994). *A Course in Game Theory*. Massachusetts: MIT press.
38. Belnap, N., & Perloff, M. (1992). The way of the agent. *Studia Logica, 51*, 463–484.
39. Widerker, D. (2017). *Moral Responsibility and Alternative Possibilities: Essays on the Importance of Alternative Possibilities*. London: Routledge.
40. Yazdanpanah, V., Dastani, M., Alechina, N., Logan, B., & Jamroga, W. (2019). Strategic responsibility under imperfect information. In: *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS-19)*, (pp. 592–600).
41. Baier, C., Funke, F., & Majumdar, R. (2021). A game-theoretic account of responsibility allocation. In: *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI-21)*, (pp. 1773–1779).
42. Lorini, E., & Schwarzentruber, F. (2011). A logic for reasoning about counterfactual emotions. *Artificial Intelligence, 175*(3), 814–847.
43. Naumov, P., & Tao, J. (2019). Blameworthiness in strategic games. In: *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI-19)*, (pp. 3011–3018)
44. Naumov, P., & Tao, J. (2020). An epistemic logic of blameworthiness. *Artificial Intelligence, 283*, 103269.
45. Chellas, B. F. (1969). *The Logical Form of Imperatives*. California: Stanford University.
46. Belnap, N., & Perloff, M. (1990). Seeing to it that: a canonical form for agentives. In: *Knowledge Representation and Defeasible Reasoning*, (pp. 167–190). Springer, Dordrecht.
47. Horty, J. F. (2001). *Agency and Deontic Logic*. Oxford: Oxford University Press.
48. Horty, J., & Pacuit, E. (2017). Action types in STIT semantics. *The Review of Symbolic Logic, 10*(4), 617–637.
49. Perloff, M. (1991). STIT and the language of agency. *Synthese, 86*, 379–408.
50. Horty, J. F., & Belnap, N. (1995). The deliberative STIT: a study of action, omission, ability, and obligation. *Journal of Philosophical Logic, 24*(6), 583–644.
51. Xu, M. (1998). Axioms for deliberative STIT. *Journal of Philosophical Logic., 27*(5), 505–552.
52. Balbiani, P., Herzig, A., & Troquard, N. (2008). Alternative axiomatics and complexity of deliberative stit theories. *Journal of Philosophical Logic, 37*(4), 387–406.
53. Olkhovikov, G. K., & Wansing, H. (2019). Inference as doxastic agency. Part I: The basics of justification STIT logic. *Studia Logica, 107*(1), 167–194.
54. Lorini, E., Longin, D., & Mayor, E. (2014). A logical analysis of responsibility attribution: emotions, individuals and collectives. *Journal of Logic and Computation, 24*(6), 1313–1339.
55. Abarca, A. I. R., & Broersen, J. M. (2022). A STIT logic of responsibility. In: *Proceeding of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS-22)*, (pp. 1717–1719).
56. Broersen, J., Herzig, A., & Troquard, N. (2006). A STIT-extension of ATL. In: *European Workshop on Logics in Artificial Intelligence*, (pp. 69–81). Springer.
57. 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*, (vol. 3, pp. 484–487). IEEE.
58. Broersen, J., & Herzig, A. (2015). Using STIT theory to talk about strategies. *Models of Strategic Reasoning: Logics, Games, and Communities*, (pp. 137–173).
59. Edwards, J. (2021). Theories of criminal law. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* California: Metaphysics Research Lab, Stanford University.
60. von Neumann, J., & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press, Princeton.

61. Ichikawa, J. J., & Steup, M. (2024). The analysis of knowledge. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy.* California: Metaphysics Research Lab, Stanford University.
62. Cui, X., & Naumov, P. (2024). Responsibility in infinite games. *Notre Dame Journal of Formal Logic, 1*(1), 1–16.
63. Knight, S., Naumov, P., Shi, Q., & Suntharraj, V. (2022). Truth set algebra: a new way to prove undefinability. arXiv:2208.04422
64. Deuser, K., Jiang, J., Naumov, P., & Zhang, W. (2024). A dynamic logic of data-informed knowledge. *Journal of Philosophical Logic,* 1–37.
65. 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.
66. Himmelreich, J. (2019). Responsibility for killer robots. *Ethical Theory and Moral Practice, 22*(3), 731–747.
67. Hindriks, F., & Veluwenkamp, H. (2023). The risks of autonomous machines: from responsibility gaps to control gaps. *Synthese, 201*(1), 21.
68. Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62–77.
69. Robillard, M. (2018). No such thing as killer robots. *Journal of Applied Philosophy, 35*(4), 705–717.
70. Oimann, A. -K. (2023). Why command responsibility may (not) be a solution to address responsibility gaps in laws. *Criminal Law and Philosophy,* 1–27.
71. Xu, M. (1998). Axioms for deliberative stit. *Journal of Philosophical Logic, 27*, 505–552.
72. Bullock, S. (2015). Executive Order Granting Clemency to Barry Allan Beach. https://formergovernors.mt.gov/bullock/docs/2015EOs/EO_19_2015_Beach.pdf. Accessed 14 May 2023
73. Montana Innocence Project (2023) Never, ever, ever give up: Barry Beach's resilient fight for freedom. https://mtinnocenceproject.org/barry-beach/. Accessed 26 May 2023
74. Sverdlik, S. (1987). Collective responsibility. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 51*(1), 61–76.
75. Mellema, G. (2006). Collective responsibility and contributing to an outcome. *Criminal Justice Ethics, 25*(2), 17–22.
76. List, C. (2021). Group agency and artificial intelligence. *Philosophy & technology, 34*(4), 1213–1242.
77. Miller, S. (2001). *Collective responsibility. Public Affairs Quarterly, 15*(1), 65–82.
78. Narveson, J. (2002). Collective responsibility. *The Journal of Ethics, 6*, 179–198.
79. Collins, S. (2019). Collective responsibility gaps. *Journal of Business Ethics, 154*, 943–954.
80. Taylor, I. (2024). Collective responsibility and artificial intelligence. *Philosophy & Technology, 37*(1), 27.