Responsibility Gap in Collective Decision Making

Pavel Naumov[1] , **Jia Tao**[2]
[1]University of Southampton, United Kingdom
[2]Lafayette College, United States
p.naumov@soton.ac.uk, taoj@lafayette.edu

Abstract

The responsibility gap is a set of outcomes of a collective decision-making mechanism in which no single agent is individually responsible. In general, when designing a decision-making process, it is desirable to minimise the gap.

The paper proposes a concept of an elected dictatorship. It shows that, in a perfect information setting, the gap is empty if and only if the mechanism is an elected dictatorship. It also proves that in an imperfect information setting, the class of gap-free mechanisms is positioned strictly between two variations of the class of elected dictatorships.

1 Introduction

AI agents are involved in making significant decisions in our everyday lives – from driving autonomous vehicles and investing in stock to estimating (in the role of war robots) potential civilian casualties. For such decisions to be socially acceptable, there should be at least one agent responsible for the outcome of the decision. That is, the decision-making mechanism should be designed without responsibility gaps.

The term *responsibility gap* (or responsibility void) is used in the literature in two distinct but related ways. First, it refers to situations where an agent who would normally be held responsible lacks *moral agency*–for example, minors, animals, and often AI systems [Matthias, 2004; Champagne and Tonkens, 2015; Burton *et al.*, 2020; Coeckelbergh, 2020; Gunkel, 2020; Santoni de Sio and Mecacci, 2021; Tigard, 2021; Königs, 2022; Oimann, 2023; Hindriks and Veluwenkamp, 2023]. Second, it describes cases where the design of a collective decision-making mechanism is such that no *single* agent (artificial or otherwise) can be held accountable for the outcome of the group decision [Braham and van Hees, 2011; Duijf, 2018; List, 2021; Duijf, 2022; Dastani and Yazdanpanah, 2023; Shi and Naumov, 2025].

In this paper, we investigate the properties of responsibility gaps–specifically in the second sense–using the concept of an elected dictatorship that we introduce. We demonstrate that in the perfect information setting, the responsibility gap is empty if and only if the mechanism constitutes an elected dictatorship. In the case of imperfect information, we further show that the class of gap-free mechanisms lies strictly between two variants of elected dictatorships. After the statement of Theorem 1, we compare our findings with two closely related results from the literature.

2 Decision-making Mechanisms

An example of a collective decision-making mechanism is the Two-person Rule used to launch American Minuteman II intercontinental ballistic missiles with nuclear warheads. Only the President of the United States can authorise the launch of the missiles. Once the President issues a launch order, the crew on the missile launch site must strap to their chairs (in case of a nuclear attack on the launch facility). Then, two on-duty officers must *simultaneously* turn their keys to activate the launch. No single officer can turn both keys because they are 12 feet apart [United States Air Force, 2024].

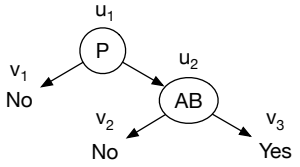


Figure 1: Two-person Rule mechanism.

Figure 1 depicts this mechanism as a tree. In this figure, the President (P) can unilaterally decide not to launch the missiles and, thus, transition the decision-making process from node u_1 to node v_1. Alternatively, the President can transition the mechanism to node u_2. In that node, officers A and B must simultaneously turn the keys in order for the decision path to end in node v_3 where the missiles are automatically launched[1]. If either of them does not turn the key, the decision-making process transitions from node u_2 to node v_2 and the missiles are not launched.

The next definition generalises the mechanism depicted in Figure 1.

Definition 1. *A tuple* $(V, E, \mathcal{A}, \Delta, \tau, \ell)$ *is a decision-making mechanism, where*

[1]Technically, for missiles to leave the silos, another pair of keys must be turned at another launch control facility [United States Air Force, 2024]. However, this does not change the responsibility analysis in this paper.

1. *(V, E) is a rooted directed tree; by L and D we denote the set of all leaf and decision (non-leaf) nodes of this tree, respectively. For each decision node $v \in D$, by Ch_v we denote the set $\{u \in V \mid (v, u) \in E\}$ of children of node v,*
2. *$\mathcal{A}$ is a set of "agents",*
3. *Δ_v^a is a nonempty set of "actions" available to an agent $a \in \mathcal{A}$ at a decision node $v \in D$,*
4. *$\tau_v : \prod_{a \in \mathcal{A}} \Delta_v^a \to Ch_v$ is a choice function for each decision node $v \in D$,*
5. *$\ell : L \to \{Yes, No\}$ is a labelling function that maps leaf nodes into "outcomes".*

Although trees are usually assumed to be finite, the results of this paper also hold for a rooted directed tree (V, E) of infinite width (but not infinite depth).

In the case of the mechanism in Figure 1, nodes u_1 and u_2 are the decision nodes and nodes v_1, v_2, and v_3 are the leaf nodes. Relation E is represented by the directed edges in the figure. In this example, the set of agents consists of the President P and officers A and B. Each of the agents has two actions: "go left" (Left) and "go right" (Right).

The choice function τ_v determines which node the decision process transitions to based on the actions of agents taken at node v. Intuitively, we assume that only the President decides at node u_1 in Figure 1 and only the two officers contribute to the decision at node u_2. This can be captured in the more general setting of Definition 1 by assuming that all three agents act at each decision node, but some actions might not influence the decision at all. For example, function τ_{u_1} formally takes a tuple (representing actions of agents P, A, and B). However, the value of this function is completely determined by the action of agent P alone. Similarly, the value of τ_{u_2} is completely determined by the actions of agents A and B.

Labelling function ℓ specifies the outcome of the decision-making process at each of the leaf nodes. Note that in this paper, we only consider the mechanisms that make binary (*Yes/No*) decisions. We briefly discuss a more general class of mechanisms in the conclusion.

Note that each element δ of the set $\Pi_{a \in \mathcal{A}} \Delta_v^a$ specifies a possible combination of actions of all agents at a decision node v. We refer to such a combination δ as an *action profile* at node v. By δ_a we denote the action of agent $a \in \mathcal{A}$ under the profile δ. Intuitively, by $Next_d^a(v)$ we denote the set of all children of node v to which the decision-making process can transition from node v if agent a chooses action d.

Definition 2. $Next_d^a(v) = \{\tau_v(\delta) \mid \delta_a = d\}$.

In the case of our running example, $Next_{Left}^A(u_2) = \{v_2\}$ and $Next_{Right}^A(u_2) = \{v_2, v_3\}$.

3 Counterfactual Responsibility

Imagine now a situation, when the President decides to authorise a nuclear strike, the two officers turn the keys, and half of the world is destroyed. Who is responsible for this? The notion of responsibility has been extensively studied in philosophy and law. In philosophy, one of the most commonly discussed approaches [Widerker and McKenna, 2003] is to define responsibility based on Frankfurt's principle of alternative possibilities: "*... a person is morally responsible for what he has done only if he could have done otherwise*" [Frankfurt, 1969]. In recent works in AI, "could have done otherwise" has been interpreted as having a *strategy*, at some point during the decision process, to prevent the outcome [Yazdanpanah *et al.*, 2019; Naumov and Tao, 2019; Naumov and Tao, 2020a; Baier *et al.*, 2021; Shi and Naumov, 2025]. In this paper, we use the term *counterfactual responsibility* to refer to the definition of responsibility based on Frankfurt's principle. We often omit "counterfactual" because this is the only type of responsibility that we consider in this work.

In our example, at the leaf node v_3, all three agents are counterfactually responsible for the decision to launch the missiles because each of them has had a strategy to prevent it. The President could have chosen not to authorise a nuclear strike by transitioning the decision-making process from node u_1 to node v_1. Each officer could have chosen not to turn the key, unilaterally transitioning the process from node u_2 into node v_2.

Because counterfactual responsibility is defined through having a strategy, before formally defining responsibility in an arbitrary decision-making mechanism, we need to define what we mean by "having a strategy" to prevent an outcome. Since our decision mechanisms have only two outcomes, *Yes* and *No*, preventing one of them is equivalent to achieving the other. Below, we use backward induction to formally define the set $win_a(o)$ of all nodes from which an agent a has a strategy to achieve an outcome o.

Definition 3. *For any outcome $o \in \{Yes, No\}$ and any agent $a \in \mathcal{A}$, let set $win_a(o)$ be the smallest subset of V such that*

1. *$\ell^{-1}(o) \subseteq win_a(o)$,*
2. *if $Next_d^a(v) \subseteq win_a(o)$, then $v \in win_a(o)$, for each decision node $v \in D$ and each action $d \in \Delta_v^a$.*

For example, for the decision mechanism depicted in Figure 1, we have $win_P(No) = \{u_1, v_1, v_2\}$, $win_P(Yes) = \{v_3\}$, $win_A(No) = \{u_1, v_1, u_2, v_2\}$, $win_A(Yes) = \{v_3\}$.

For any outcome $o \in \{Yes, No\}$, by $\overline{o}$ we mean the other outcome. For example, $\overline{Yes} = No$.

Definition 4. *A **decision path** is any sequence of nodes $v_1, \dots, v_k$ such that $k \geq 1$ and $v_{i+1} \in Ch_{v_i}$ for each $i < k$.*

The next definition formally captures Frankfurt's principle of alternative possibilities.

Definition 5. *An agent $a \in \mathcal{A}$ is **responsible** at a leaf node $v \in L$ if there is a decision node $u \in D$ on the decision path from the root to leaf v such that $u \in win_a(\overline{\ell(v)})$.*

Let us now consider a situation when the United States is attacked by an enemy using nuclear weapons, but the President fails to authorise a retaliatory strike. The decision-making process terminates in node v_1, see Figure 1. Note that the President does not have a strategy to guarantee the strike because the President cannot guarantee that the two keys will be simultaneously turned by officers A and B. Thus, the President is not *counterfactually* responsible in such a situation[2].

[2]In outcome v_1 the President is responsible for "seeing-to-it" that

Of course, in this situation, the two officers are not counterfactually responsible either. Formally, by Definition 3,

$$u_1, v_1 \notin win_P(\textit{Yes}) \cup win_A(\textit{Yes}) \cup win_B(\textit{Yes}).$$

Hence, in such a situation, none of the agents is responsible for the lack of a retaliatory strike. In other words, there is a *responsibility gap*.

The responsibility gap also exists in node v_2, where, after the President authorises the launch, at least one of the officers decides not to turn the key. This is because even after the President's authorisation, neither officer has a strategy to guarantee the launch. After all, the other officer can always decide not to turn the key. Formally,

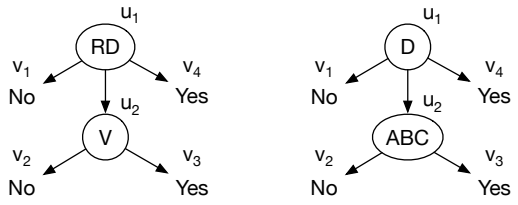
$$u_1, u_2, v_2 \notin win_P(\textit{Yes}) \cup win_A(\textit{Yes}) \cup win_B(\textit{Yes}).$$

Definition 6. *A mechanism is **gap-free** if, for each leaf node, there is at least one agent responsible at this node.*

The mechanism depicted in Figure 1 is *not* gap-free because no agent is responsible at leaf nodes v_1 and v_2.

4 Elected Dictatorship

Next, let us turn our attention to a completely different setting inspired by Article I of the US Constitution: *"The Vice President of the United States shall be President of the Senate, but shall have no Vote, unless they be equally divided"*. To make the example more manageable, let us suppose that the Senate contains just two senators: a Republican (R) and a Democrat (D). If their votes (by a paper ballot) agree, the decision is made. Otherwise, the Vice President breaks the tie, see Figure 2a.



(a) US Senate mechanism. (b) Academic mechanism.

Figure 2: (a) There is a responsibility gap at nodes v_1 and v_4. The Vice President is a dictator at node u_2. (b) The mechanism is gap-free. The Dean is a dictator at root node u_1.

In this setting, the responsibility gap exists in node v_1 (outcome "*No*") and node v_4 (outcome "*Yes*"). This is because if the process terminates in either of these nodes, then none of the three agents, at any point during the decision-making process, has a unilateral strategy to prevent the outcome:

$$u_1, v_1 \notin win_R(\textit{Yes}) \cup win_D(\textit{Yes}) \cup win_V(\textit{Yes}),$$
$$u_1, v_4 \notin win_R(\textit{No}) \cup win_D(\textit{No}) \cup win_V(\textit{No}).$$

retaliatory strike does not take place. The "seeing-to-it" form of responsibility (see [Shi and Naumov, 2025] for an overview) is different from the counterfactual responsibility that we consider in this paper.

However, the Vice President is counterfactually responsible for the outcome in nodes v_2 and v_3 because, if the decision-making process reaches either node, then the Vice President has had a chance to prevent the outcome corresponding to the node. In fact, at node u_2, the Vice President simultaneously had a strategy to guarantee either of the two possible outcomes: $u_2 \in win_V(\textit{Yes})$ and $u_2 \in win_V(\textit{No})$. We say that the Vice President is a *dictator* at node u_2.

Definition 7. *An agent $a \in \mathcal{A}$ is a **dictator** at a decision node $v \in D$ if $v \in win_a(\textit{Yes})$ and $v \in win_a(\textit{No})$.*

There is no dictator at any of the nodes in the Two-person Rule mechanism depicted in Figure 1.

In this paper, we investigate the connection between responsibility gaps and the presence of dictators in a decision mechanism. The example depicted in Figure 2a shows that the existence of a single dictator is not enough to guarantee that the mechanism is gap-free. However, as we are about to see, the existence of a single dictator condition can be strengthened to provide such a guarantee.

Towards this goal, let us consider one more example. It seems to be a common pattern in academia that administrators prefer to avoid making unpopular decisions by delegating the decision-making to a committee. We capture this situation in the mechanism depicted in Figure 2b. Here, the Dean (D) can either decide on *Yes/No* or delegate the decision to a three-member committee consisting of academic staff members A, B, and C. The committee makes the decision by a majority vote using a paper ballot.

In such a setting, each of the three committee members is never *individually* responsible for the outcome because none of them at any moment has an individual strategy that guarantees any of the two outcomes. At the same time, the Dean is not only responsible for the decision in all four leaf nodes, but the Dean is also a dictator at the root node u_1. In particular, this means that there is a dictator at a node on each root-to-leaf decision path of the mechanism.

Definition 8. *A mechanism is an **elected dictatorship** if there is a dictator at one of the decision nodes of each root-to-leaf decision path.*

The mechanism depicted in Figure 2b is an extreme example of an elected dictatorship where there is a single dictator at the root node. More generally, different agents might be dictators along different root-to-leaf decision paths of an elected dictatorship.

It turns out that being an elected dictatorship is not just a sufficient condition for being gap-free, but these two conditions are equivalent. See the theorem below.

Theorem 1. *A mechanism is a gap-free mechanism iff it is an elected dictatorship.*

The proof of the above theorem, as well as the missing proofs of other results, can be found in the full version of this paper [Naumov and Tao, 2025]. As we show there, Theorem 1 follows from more general results about games with imperfect information that we establish later in this paper. This theorem gives a complete characterisation of the "gap-freeness" for the class of decision-making mechanisms specified in Definition 1. There are two previous related works

that considered this property for a much more narrow class of "discursive dilemma" mechanisms. In terms of Definition 1, a discursive dilemma mechanism is a *single-node* mechanism with function τ having a very special "criteria-based" form [List, 2006]. Duijf and van De Putte [2022] considered an alternative best-effort-based definition of responsibility and gave a complete characterisation of gap-free mechanisms in discursive dilemma mechanisms. Their characterisation does not refer to dictatorship. Braham and van Hees [2018] considered probabilistic discursive dilemma mechanisms and another variation of the definition of responsibility. They proved that if a gap-free discursive dilemma mechanism does not have what they call "fragmentation", then the mechanism must be a dictatorship.

5 Mechanisms with Imperfect Information

In many decision-making mechanisms, some agents might not have complete information about the actions already taken. An example of such a mechanism is the Drawing Straws[3]. Figure 3 shows a version of this mechanism in which Alice (A) holds both a short and a long straw between her fingers, without revealing which straw is which. Bob (B) picks either left (action 0) or right (action 1) straw. The outcome of the decision-making is determined by whether he has chosen a short or a long straw. The dashed line labelled with B in the figure represents the fact that Bob cannot distinguish nodes u_2 and u_3 at the moment he chooses an action at those nodes.

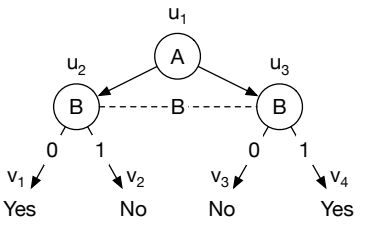


Figure 3: Drawing Straws mechanism.

To handle the decision-making processes like the one in Figure 3, we need a more general notion of a decision-making mechanism in which the agents cannot distinguish some of the decision nodes.

Definition 9. *A decision-making mechanism with imperfect information is a tuple* $(V, E, \mathcal{A}, \Delta, \tau, \ell, \sim)$ *where*

1. $(V, E, \mathcal{A}, \Delta, \tau, \ell)$ *is a decision-making mechanism,*
2. $\sim_a$ *is an* ***indistinguishability*** *equivalence relation on the set* D *of decision nodes for each agent* $a \in \mathcal{A}$ *such that if* $u \sim_a v$*, then* $\Delta^a_u = \Delta^a_v$.

Note that the "if $u \sim_a v$, then $\Delta^a_u = \Delta^a_v$" requirement of item 2 above specifies that each agent has the same available actions in all indistinguishable nodes. In other words, each agent *knows* the actions available to her at the current node.

[3]In 2017, this mechanism was used to decide who gets a seat in the Northumberland County Council (England) after votes were evenly divided [Elgot, 2017].

The set of decision-making mechanisms as specified in Definition 1 consists of quintuples, while the set of mechanisms with imperfect information in Definition 9 consists of sextuples. Using the terminology of *object-oriented programming* languages, we can say that the class of mechanisms with imperfect information is an *extension* of the class of mechanisms from Definition 1. Indeed, to treat a mechanism with imperfect information as a mechanism, we just need to ignore the equivalence relations. From this point of view, Definition 2 through Definition 8 are still applicable to the mechanisms with imperfect information. Using the object-oriented terminology, one can say that the class of mechanisms with imperfect information *inherits* the notions specified in these definitions from the generic class of mechanisms. We adopt such a viewpoint in this paper.

6 Epistemic Responsibility

Let us go back to the Drawing Straws mechanism depicted in Figure 3. Suppose that Alice positions long and short straws in such a way that the mechanism transitions from node u_1 to node u_2. Does Bob have a strategy at node u_2 to guarantee the outcome *No*? We would say that he does (it is action 1). In fact, it is easy to see that $u_2 \in win_B(No)$ by Definition 3. Thus, by Definition 5, Bob is counterfactually responsible at leaf node v_1. This observation, however, is not intuitively acceptable: how can Bob be blamed for pulling, say, a long straw if he did not know which of the two straws is long and which is short? This is the reason why in the literature it has been suggested that in order for an agent to be counterfactually responsible in an imperfect information setting, the agent should not only have a strategy to prevent the outcome but also should know what this strategy is [Yazdanpanah *et al.*, 2019; Naumov and Tao, 2020b].

To define what "to know the strategy" formally means in our setting is a non-trivial task. In the literature, uniform or "know-how" [Fervari *et al.*, 2017; Naumov and Tao, 2018] strategies are usually defined as functions that assign the same actions to all indistinguishable nodes. Following this approach, we can adjust Definition 3 for the imperfect information setting as shown below. By $[v]_a$ we mean the equivalence class of node v with respect to the equivalence relation $\sim_a$.

Definition 10. *For any outcome* $o \in \{Yes, No\}$ *and any agent* $a \in \mathcal{A}$*, let set* $uwin_a(o)$ *be the smallest subset of* V *such that, for each node* $v \in V$*,*

1. $\ell^{-1}(o) \subseteq uwin_a(o)$*,*
2. *if* $Next^a_d([v]_a) \subseteq uwin_a(o)$*, then* $v \in uwin_a(o)$*, for each decision node* $v \in D$ *and each action* $d \in \Delta^a_v$.

To define the notion of counterfactual responsibility in decision-making mechanisms with imperfect information, one can consider replacing the set $win_a(\overline{\ell(v)})$ with the set $uwin_a(\overline{\ell(v)})$ in Definition 5. Unfortunately, this does *not* capture our intuition of what is responsibility in an imperfect information setting. Indeed, consider the decision mechanism with imperfect information depicted in Figure 4.

This mechanism has a single agent A (we don't need additional agents to explain the issue). In each decision node, this

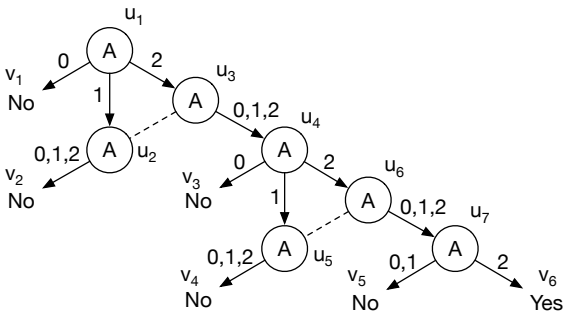


Figure 4: A single-agent decision-making mechanism with imperfect information. Dashed lines represent the relation $\sim_A$.

agent has the same set of actions $\{0, 1, 2\}$. The outcomes of these actions are shown in the figure.

Note that $v_6 \in uwin_A(\textit{Yes})$ by item 1 of Definition 10. Also, $u_7 \in uwin_A(\textit{Yes})$ by item 2 of Definition 10 (observe that $[u_7]_A = \{u_7\}$ and take $d = 2$). At the same time, $u_6 \notin uwin_A(\textit{Yes})$. Indeed, $[u_6]_A = \{u_5, u_6\}$ and there is no action d such that $Next^a_d([u_6]_a) \subseteq uwin_a(\textit{Yes})$. As a result, again by Definition 10,

$$u_4 \notin uwin_A(\textit{Yes}). \tag{1}$$

Next, consider a case when the decision process terminates at leaf node v_3. The outcome of this process is *No*. Could agent A be held counterfactually responsible for this? Intuitively, yes! To reach node v_3, the decision path must go through node u_4. Agent A can distinguish node u_4 from all other nodes in the mechanism, see Figure 4. Thus, while the process is at node u_4, agent A knows that the process is at this node. Hence, while at node u_4, agent A knows that if she chooses action 2, then she will become "confused" (mechanism will transition to node u_6 where she will not know how to achieve outcome *Yes*). However, in node u_4, agent A knows that no matter what she does while being "confused" (in node u_6), the process will come to a node (in our case u_7) where she will wake up from the confusion and will know how to achieve the outcome *Yes*. Having all this information in node u_4, agent A, we believe, "knows" how to achieve outcome *Yes* – take a deep breath and put herself in the state of confusion by choosing action 2. Thus, we think, she should be held counterfactually responsible for the outcome *No* in leaf node v_3. To achieve this, we need to modify Definition 10 in such a way that statement (1) is no longer true.

Of course, one might argue that the issue that we described in the previous paragraph only exists because we allow decision-making mechanisms in which agents can get "confused". If we assume that the agents have perfect recall, then the situation depicted in Figure 4 will never happen. Specifically, if agent A remembers what action she took in node u_4, then she will always be able to distinguish node u_5 from node u_6. Thus, at node u_6 she would still know how to achieve the outcome *Yes*.

We agree that Definition 10 *seems* to work for agents with perfect recalls and that most commonly used decision-making mechanisms do not force agents to forget what they know. However, there are examples of decision making mechanisms that do so. For instance, a new employee selection mechanism might ask the selection committee to ignore information not shown in the application materials. A judge in court might instruct the jury to ignore certain evidence or a witness testimony. To state our results in the most general form that covers such mechanisms, in this paper we modify Definition 10. Our revised definition will be able to handle multiple "confusions/memory losses" during the decision-making process. For example, consider leaf node v_1 in the same mechanism depicted in Figure 4. We believe agent A is counterfactually responsible for the outcome *No* at this node. Indeed, at node u_1 agent A knows how to achieve outcome *Yes* – by putting herself into the state of confusion twice: first, by taking action 2 at node u_1 and then by taking the same action 2 at node u_4. Our replacement for Definition 10 is Definition 11 stated below.

Definition 11. *For any outcome $o \in \{\textit{Yes}, \textit{No}\}$ and any agent $a \in \mathcal{A}$, let set $ewin_a(o)$ be the smallest subset of V such that, for each node $v \in V$,*

1. *$\ell^{-1}(o) \subseteq ewin_a(o)$,*
2. *if $Next^a_d([v]_a) \subseteq ewin_a(o)$, then $v \in ewin_a(o)$, for each decision node $v \in D$ and each action $d \in \Delta^a_v$,*
3. *if $\bigcup_{d \in \Delta^a_v} Next^a_d(v) \subseteq ewin_a(o)$, then $v \in ewin_a(o)$, for each decision node $v \in D$.*

Note that Definition 11 adds one extra recursive case (item 3) to Definition 10. This item states that the agent does not need to know how to act in the current node if any possible action in this node leads to the set $ewin_a(o)$. Because Definition 11 adds an *extra* case, $uwin_a(o) \subseteq ewin_a(o)$ for each agent $a \in \mathcal{A}$ and each outcome $o \in \{\textit{Yes}, \textit{No}\}$.

In Figure 4, for example, $v_6 \in ewin_A(\textit{Yes})$ by item 1 of Definition 11. Then, $u_6 \in ewin_A(\textit{Yes})$ by item 3 of Definition 11. Hence, $u_4 \in ewin_A(\textit{Yes})$ by item 2 of Definition 11 with $d = 2$.

As we discussed after Definition 9, each mechanism with imperfect information can be viewed as a "mechanism" under Definition 1. Thus, for the mechanisms with imperfect information, in addition to set $ewin_a(o)$, one can also consider the set $win_a(o)$ as specified in Definition 3. For example, $ewin_B(\textit{Yes}) = \{v_1, v_4\}$ and $win_B(\textit{Yes}) = \{v_1, v_4, u_2, u_3\}$ for the Drawing Straw mechanism in Figure 3.

The next lemma connects these two sets for an arbitrary mechanism with imperfect information.

Lemma 1. *$ewin_a(o) \subseteq win_a(o)$, for each agent $a \in \mathcal{A}$ and each outcome $o \in \{\textit{Yes}, \textit{No}\}$.*

We are now ready to define what it means to be counterfactually responsible in decision-making mechanisms with imperfect information. Our definition below simply replaces the set $win_a(\overline{\ell(v)})$ with the set $ewin_a(\overline{\ell(v)})$ in Definition 5. Since we consider mechanisms with imperfect information to be a "subclass" (in the sense of object-oriented programming) of mechanisms, the notion "responsible", as specified in Definition 5 is still technically defined for the mechanisms with imperfect information. To avoid confusion, we use the term "epistemically responsible" in the definition below. However,

it is important to remember that "epistemically responsible" is *the* proper definition of being counterfactually responsible in an imperfect information setting. It is the notion that properly captures our intuition about responsibility.

Definition 12. *An agent $a \in \mathcal{A}$ is **epistemically responsible** at a leaf node $v \in L$ if there is a decision node $u \in D$ on the decision path from the root to the leaf node v such that $u \in ewin_a(\overline{\ell(v)})$.*

For example, agent A is epistemically responsible at all leaf nodes of the mechanism depicted in Figure 4.

Intuitively, by an "epistemic gap" we mean the set of all leaf nodes in which no agent is epistemically responsible.

Definition 13. *A mechanism is **epistemic-gap-free** if, for each leaf node, there is at least one agent epistemically responsible at this node.*

7 Elected Epistemic Dictatorship

The notions of "dictator at a node" and "elected dictatorship", as specified in Definition 7 and Definition 8, have their epistemic counterparts based on function $ewin$ instead of win.

Definition 14. *An agent $a \in \mathcal{A}$ is an **epistemic dictator** at a node v if $v \in ewin_a(Yes)$ and $v \in ewin_a(No)$.*

Definition 15. *A mechanism is an **elected epistemic dictatorship** if, for each root-to-leaf decision path, there is an epistemic dictator at a node on this path.*

The next theorem shows that the epistemic version of the right-to-left part of Theorem 1 is true.

Theorem 2. *Any elected epistemic dictatorship is epistemic-gap-free.*

Proof. Consider any root-to-leaf path $v_1, \ldots, v_n$. By Definition 13, it suffices to show that some agent is epistemically responsible at leaf node v_n. By the assumption of the theorem that the mechanism is an elected epistemic dictatorship and Definition 15, there is $i < n$ and an agent $a \in \mathcal{A}$ who is an epistemic dictator at decision node v_i. Hence, $v_i \in ewin_a(Yes)$ and $v_i \in ewin_a(No)$ by Definition 14. Thus, $v_i \in ewin_a(\overline{\ell(v_n)})$. Therefore, agent a is epistemically responsible at leaf node v_n by Definition 12. □

Perhaps surprisingly, the epistemic version of the other direction of Theorem 1 is false. To prove this, consider the decision-making mechanism with imperfect information M depicted in Figure 5.

Lemma 2. *Mechanism M with imperfect information is epistemic-gap-free.*

Note that agent B is a dictator at node u_2 because this agent can use action 0 to transition the decision-making process from node u_2 to node $u_5 \in win_B(No)$ and this agent can use action 1 to transition the decision-making process from node u_2 to node $v_1 \in win_B(Yes)$. However, as we show in the proof of the next lemma, agent B is not an *epistemic* dictator at node u_2.

Lemma 3. *Mechanism M with imperfect information is not an epistemic elected dictatorship.*

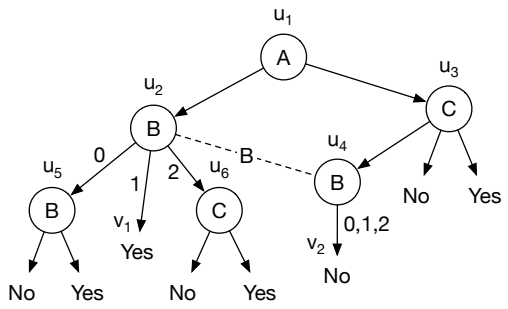


Figure 5: Mechanism M with imperfect information. The names of actions are only shown for the nodes in which the acting agent does not have complete information about the current node.

Together, Lemma 2 and Lemma 3 show that mechanism M provides a counterexample to the converse of Theorem 2.

In conclusion, observe that mechanism M, unlike the mechanism shown in Figure 4, does not make any agent to forget something that she knew before by transitioning the agent into a "confused" state. Thus, this mechanism is suitable for agents with perfect recall. Additionally, observe that the proof of Theorem 2 does not rely on the specific definition of the function $ewin$. For example, if the notion of epistemic dictator and epistemic responsibility are defined using function $uwin$ instead of $ewin$, the proof of Theorem 2 remains valid. Furthermore, the counterexample given by mechanism M also works for many variations of $ewin$ definition. In particular, it is easy to see that the proofs of Lemma 2 and Lemma 3 remain valid if the notion of epistemic dictator and epistemic responsibility are defined using function $uwin$ instead of $ewin$. These observations show that the results of this section appear to be very general and are not artifacts of the specifics of Definition 11.

8 Elected Semi-epistemic Dictatorship

In Theorem 2, we have shown that, for the mechanisms with imperfect information, the set of elected epistemic dictatorships is a subset of the set of epistemic-gap-free mechanisms. We used the mechanism M to show that the former set is a *proper* subset of the latter. Given that Theorem 1 shows that these sets are equal for the mechanism with perfect information, it is natural to ask if the notion of elected epistemic dictatorship could be made *weaker* so that the modified set of elected epistemic dictatorships includes the set of all epistemic-gap-free mechanisms. In this section, we propose such a modification: elected semi-epistemic dictatorship.

Definition 16. *An agent $a \in \mathcal{A}$ is a **semi-epistemic dictator** at a node v if there exists an outcome $o \in \{Yes, No\}$ such that $v \in ewin_a(o)$ and $v \in win_a(\overline{o})$.*

As an example, note that agent B is a semi-epistemic dictator at node u_2 of the mechanism M depicted in Figure 5. Indeed, it is easy to verify that $u_2 \in ewin_B(No)$ and $u_2 \in win_B(Yes)$.

Definition 17. *A mechanism with imperfect information is an **elected semi-epistemic dictatorship** if, for each root-to-leaf*

decision path, there is a semi-epistemic dictator at a node on this path.

An example of an elected semi-epistemic dictatorship is mechanism M depicted in Figure 5. Indeed, agents B and C are semi-epistemic dictators at nodes u_2 and u_3, respectively. In fact, C is not just a semi-epistemic dictator, but also an epistemic dictator at node u_3.

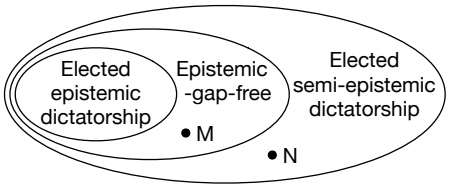


Figure 6: Three sets of decision-making mechanisms with imperfect information. Mechanisms M and N are shown in Figure 5 and Figure 7, respectively.

In this section, we prove that epistemic-gap-free mechanisms form a subset of elected semi-epistemic dictatorships, see Figure 6. We prove this result in Theorem 3.

Lemma 4. *For any root-to-node decision path $v_1, \dots, v_k$ of an epistemic-gap-free mechanism, if $v_k \in ewin_a(o)$ and $v_k \notin win_a(\overline{o})$, then there exists $i < k$ and $b \in \mathcal{A}$ such that $v_i \in ewin_b(\overline{o})$.*

Theorem 3. *Any epistemic-gap-free mechanism with imperfect information is an elected semi-epistemic dictatorship.*

Proof. Consider any root-to-leaf decision path $v_1, \dots, v_n$. By Definition 17, it suffices to show that there is a semi-epistemic dictator at one of the decision nodes of this path. Suppose the opposite. Thus,

$$\begin{aligned} &\text{if } v_i \in ewin_a(o), \text{ then } v_i \notin win_a(\overline{o}) \\ &\text{for each } i < n, a \in \mathcal{A}, o \in \{\text{Yes}, \text{No}\}. \end{aligned} \tag{2}$$

At the same time, by Definition 13, the assumption that the mechanism is epistemic-gap-free implies that there is an agent $b \in \mathcal{A}$ epistemically responsible at leaf node v_n. Hence, by Definition 12, there is $j < n$ such that $v_j \in ewin_b(\overline{\ell(v_n)})$. Note that $v_j \in ewin_b(\overline{\ell(v_n)})$ implies $\exists a \exists o (v_j \in ewin_a(o))$. Thus, $\exists a \exists o (v_j \in ewin_a(o))$. Then, there must exist the minimal $j_{\min} \geq 1$ such that

$$\exists a \exists o (v_{j_{\min}} \in ewin_a(o)). \tag{3}$$

Hence, $v_{j_{\min}} \in ewin_{a'}(o')$ for some agent $a' \in \mathcal{A}$ and some outcome $o' \in \{\text{Yes}, \text{No}\}$. Thus, $v_{j_{\min}} \notin win_{a'}(\overline{o'})$ by statement (2). Then, by Lemma 4 and the assumption of the theorem that the mechanism is epistemic-gap-free, there exists $j' < j_{\min}$ and an agent $b' \in \mathcal{A}$ such that $v_{j'} \in ewin_{b'}(\overline{o'})$. The last statement contradicts the choice of $j_{\min}$ as the minimal one satisfying condition (3). □

In Theorem 3, we have shown that the set of epistemic-gap-free mechanisms is a subset of the set of elected semi-epistemic dictatorships. This subset is *proper*, see Figure 6. To prove this, let us modify the Drawing Straws mechanism depicted in Figure 3. Recall that this mechanism is sometimes used to determine the outcome of an election when votes are evenly divided. Suppose that outcome *Yes* means that Bob (and not Alice) becomes an elected official. We modify the Drawing Straws mechanism by giving Bob an option to give up the race and let Alice to become the elected official. This option is represented by action 2 in Figure 7 showing the modified mechanism N.

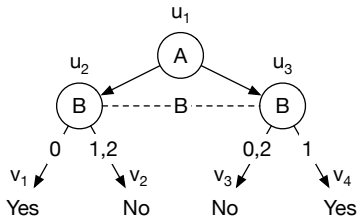


Figure 7: Mechanism N with imperfect information. The names of actions are only shown for the nodes in which the acting agent does not have complete information about the current node.

Lemma 5. *Mechanism N with imperfect information is an elected semi-epistemic dictatorship.*

Lemma 6. *Mechanism N with imperfect information is not epistemic-gap-free.*

Together, the last two lemmas show that the set of epistemic-gap-free mechanisms is a *proper* subset of the set of elected semi-epistemic dictatorships, see Figure 6.

9 Conclusion

This paper contains two main results. First, in the perfect information case, the only way to avoid a responsibility gap in a decision-making mechanism is to use an "elected dictatorship", where the agents agree on a single person who makes the decision. The converse is also true: any elected dictatorship is gap-free. Second, in the imperfect information case, the situation is more complicated: epistemic-gap-free mechanisms are "squeezed" between elected epistemic and elected semi-epistemic dictatorships. Intuitively, our results mean that to construct non-dictatorial gap-free mechanisms, one needs to consider other forms of responsibility.

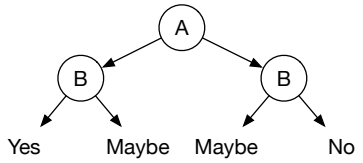


Figure 8: Gap-free mechanism with three alternatives which is not an elected dictatorship.

Note that all our results assume that the decision is binary (*Yes/No*). If the third ("*Maybe*") alternative is added, then our results are no longer true. For example, Figure 8 depicts a gap-free mechanism (agent B can prevent any specific outcome) in which none of the agents is a dictator (has a strategy to guarantee any outcome) at any of the nodes.

References

[Baier *et al.*, 2021] Christel Baier, Florian Funke, and Rupak Majumdar. A game-theoretic account of responsibility allocation. In *30th International Joint Conference on Artificial Intelligence (IJCAI-21)*, 2021.

[Braham and van Hees, 2011] Matthew Braham and Martin van Hees. Responsibility Voids. *The Philosophical Quarterly*, 61(242):6–15, 12 2011.

[Braham and van Hees, 2018] Matthew Braham and Martin van Hees. Voids or fragmentation: Moral responsibility for collective outcomes. *The Economic Journal*, 128(612):F95–F113, 2018.

[Burton *et al.*, 2020] Simon Burton, Ibrahim Habli, Tom Lawton, John McDermid, Phillip Morgan, and Zoe Porter. Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective. *Artificial Intelligence*, 279:103201, 2020.

[Champagne and Tonkens, 2015] Marc Champagne and Ryan Tonkens. Bridging the responsibility gap in automated warfare. *Philosophy & Technology*, 28:125–137, 2015.

[Coeckelbergh, 2020] Mark Coeckelbergh. Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and engineering ethics*, 26(4):2051–2068, 2020.

[Dastani and Yazdanpanah, 2023] Mehdi Dastani and Vahid Yazdanpanah. Responsibility of ai systems. *Ai & Society*, 38(2):843–852, 2023.

[Duijf and van De Putte, 2022] Hein Duijf and Frederik van De Putte. The problem of no hands: responsibility voids in collective decisions. *Social Choice and Welfare*, 58(4):753–790, 2022.

[Duijf, 2018] Hein Duijf. Responsibility voids and cooperation. *Philosophy of the social sciences*, 48(4):434–460, 2018.

[Duijf, 2022] Hein Duijf. *The logic of responsibility voids*. Springer, 2022.

[Elgot, 2017] Jessica Elgot. Lib dem and tory candidates draw straws in northumberland vote. https://www.theguardian.com/politics/2017/may/05/lib-dem-and-tory-candidates-draw-straws-in-northumberland-vote, 2017. Accessed: 2025-01-15.

[Fervari *et al.*, 2017] Raul Fervari, Andreas Herzig, Yanjun Li, and Yanjing Wang. Strategically knowing how. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 1031–1038, 2017.

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

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

[Hindriks and Veluwenkamp, 2023] Frank Hindriks and Herman Veluwenkamp. The risks of autonomous machines: from responsibility gaps to control gaps. *Synthese*, 201(1):21, 2023.

[Königs, 2022] Peter Königs. Artificial intelligence and responsibility gaps: what is the problem? *Ethics and Information Technology*, 24(3):36, 2022.

[List, 2006] Christian List. The discursive dilemma and public reason. *Ethics*, 116(2):362–402, 2006.

[List, 2021] Christian List. Group agency and artificial intelligence. *Philosophy & technology*, 34(4):1213–1242, 2021.

[Matthias, 2004] Andreas Matthias. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology*, 6:175–183, 2004.

[Naumov and Tao, 2018] Pavel Naumov and Jia Tao. Together we know how to achieve: An epistemic logic of know-how. *Artificial Intelligence*, 262:279 – 300, 2018.

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

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

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

[Naumov and Tao, 2025] Pavel Naumov and Jia Tao. Responsibility gap in collective decision making. *arXiv:2505.06312*, 2025.

[Oimann, 2023] Ann-Katrien Oimann. The responsibility gap and laws: A critical mapping of the debate. *Philosophy & Technology*, 36(1):3, 2023.

[Santoni de Sio and Mecacci, 2021] Filippo Santoni de Sio and Giulio Mecacci. Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology*, 34(4):1057–1084, 2021.

[Shi and Naumov, 2025] Qi Shi and Pavel Naumov. Responsibility in multi-step decision schemes. *Journal of Philosophical Logic*, 2025.

[Tigard, 2021] Daniel W Tigard. There is no techno-responsibility gap. *Philosophy & Technology*, 34(3):589–607, 2021.

[United States Air Force, 2024] United States Air Force. Launching missiles. https://www.nationalmuseum.af.mil/Visit/Museum-Exhibits/Fact-Sheets/Display/Article/197675/, 2024. Accessed: 2024-09-24.

[Widerker and McKenna, 2003] David Widerker and Michael McKenna, editors. *Moral responsibility and alternative possibilities: Essays on the importance of alternative possibilities*. Ashgate, Burlington, VT, 2003.

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