Higher-Order Responsibility

Junli Jiang[1], Pavel Naumov[2]

[1] Institute of Logic and Intelligence, Southwest University, China
[2] University of Southampton, United Kingdom
walk08@swu.edu.cn, p.naumov@soton.ac.uk

Abstract

In ethics, individual responsibility is often defined through Frankfurt's principle of alternative possibilities. This definition is not adequate in a group decision-making setting because it often results in the lack of a responsible party or "responsibility gap". One of the existing approaches to address this problem is to consider group responsibility. Another, recently proposed, approach is "higher-order" responsibility. The paper considers the problem of determining whether higher-order responsibility up to a given degree is sufficient to close the responsibility gap and analyses the computational complexity of this problem.

Introduction

Autonomous agents–from self-driving cars and stock traders to military robots and medical assistants–are increasingly involved in decisions that impact human lives. These decisions often result from collaborations among multiple parties in hybrid human-machine environments. Assigning individual responsibility for the outcomes of such decisions to at least one party fosters accountability and enhances trust.

Counterfactual Responsibility As a concept, responsibility has been studied in philosophy, law, and, more recently, artificial intelligence. In philosophy, the starting point of the discussion of responsibility is often the so-called principle of alternative possibilities: "*a person is morally responsible for what he has done only if he could have done otherwise*" (Frankfurt 1969). The arguments for and against this principle have been proposed (Blumenfeld 1971; Naylor 1984; Copp 1997; Widerker and McKenna 2003; Mele and Robb 2018). Even Frankfurt argued that its applicability is limited. In this paper, we use the term *counterfactual responsibility* to refer to the definition through the principle of alternative possibilities. Following recent literature in artificial intelligence (Yazdanpanah, Dastani, Jamroga, Alechina, and Logan 2019; Naumov and Tao 2019, 2020a; Baier, Funke, and Majumdar 2021; Shi 2024), we interpret "could have done" as having a strategy that would *guarantee* the prevention of the undesired outcome no matter what the other agents do.

As an example, imagine two factories, A and B, that can potentially dump a pollutant into a river inhabited by fish (Halpern 2016). Assume that it is well-known that it takes 15kg of pollutants to kill the fish. Consider the case when factories A and B have accumulated 20kg and 10kg of the pollutant, respectively. Suppose they both dump the pollutant into the river. The fish is dead. Which of the two factories is responsible for the death of the fish? Note that factory A has had a strategy ("do not dump the pollutant") that guarantees that the fish would have stayed alive no matter what the other factory does. Factory B has not had such a strategy. Thus, only factory A is counterfactually responsible for the death of the fish.

Deontic Constraint Let us now consider a different case in which factories A and B have each accumulated only 10kg of the pollutant. Thus, the fish will die only if both of the factories dump the pollutant. In such a situation, it would be acceptable for either of the factories, but not both, to dump the pollutant. We use propositional variables p_a ("pollute a") and p_b ("pollute b") to represent the actions of factories A and B, respectively. If $p_x = 1$, then factory X dumps the pollutant. If $p_x = 0$, it does not dump. We use the *deontic constraint* formula $\gamma = \neg(p_a \wedge p_b)$ to express which combinations of actions are acceptable. In our case, the deontic constraint γ states that it is not acceptable for both companies to dump the pollutant. Let us further suppose that both factories have dumped the pollutant, and the fish is dead. Who is responsible? In this setting, each of the two companies has had a strategy ("do not dump the pollutant") to save the fish. Thus, both of them are counterfactually responsible for the death of the fish.

Responsibility Gap Next, assume that each of the two factories, A and B, has accumulated 20kg of the pollutant. Thus, the fish dies if either of the factories dumps the pollutant. In this situation, the deontic constraint is $\gamma = \neg(p_a \vee p_b)$. In our previous examples, it is not important if the factories make decisions simultaneously (independently) or consequently. The situation is different in this example. If the decisions are made simultaneously, then neither of the two factories has a strategy that *guarantees* that the fish is alive. Thus, each time when either of the factories dumps the pollutant, the fish dies and neither of the factories is counterfactually responsible for the death. By an *action profile* we mean the pair (p_a, p_b) that describes the actions of both factories. By the *responsibility gap* we mean the set of all

profiles under which the deontic constraint γ is violated and none of the agents is counterfactually responsible for this. In our example, the responsibility gap is $\{(1,0),(0,1),(1,1)\}$.

The situation is different if the factories make their actions not simultaneously, but sequentially. Suppose that factory B makes its decision after factory A. Furthermore, we assume that factory B knows the choice made by factory A. Consider action profile $(0,1)$. Note that in this situation factory B knows that factory A decided not to dump the pollutant. Thus, it now has a strategy ("do not dump") to save the fish. Nevertheless, under the profile $(0,1)$, factory B still dumps the pollutant and kills the fish. Hence, factory B is now counterfactually responsible for the death of the fish. It is easy to see that, even in the case of consecutive decisions, nobody is counterfactually responsible for the death under the profiles $(1,0)$ and $(1,1)$ because under these profiles neither of the factories has a strategy that *guarantees* that the fish stays alive. As a result, if the decisions are made consecutively, then the responsibility gap is the set

$$G = \{(1,0),(1,1)\}. \tag{1}$$

Note that, in the above example, switching from simultaneous to consecutive decision-making *reduces the responsibility gap* from $\{(1,0),(0,1),(1,1)\}$ to $\{(1,0),(1,1)\}$. This is not a coincidence. Consecutive decision-making provides additional information to the parties that decide later. This information might *enable* them to prevent a violation of the deontic constraint. If they do not use this ability and the constraint is violated, then they become counterfactually responsible. In many situations, the responsibility gap is an undesirable property of a decision-making mechanism. In such situations, the introduction of an order in which the decisions are made is one of the techniques used by mechanism designers to minimise the responsibility gap. For example, road traffic laws use "yield" or "give way" signs, the time of arrival to an intersection, and the "priority to the right" rule to establish the order in which drivers must make their decisions. This eliminates the responsibility gap and, in the case of an accident, ascribes the responsibility to a specific driver. *In the rest of this paper, we only consider decision-making mechanisms where the agents make the decisions in a consecutive order.*

Second-Order Responsibility   As we have seen above, the introduction of the order in which agents act might shrink the responsibility gap, but not be able to eliminate it completely. Under any action profile that belongs to the gap, the deontic constraint is violated and nobody is counterfactually responsible for this. However, there might be an agent who is *counterfactually responsible for the gap*. We refer to this responsibility for the gap as *second-order responsibility* for the violation of the deontic constraint. In the case when both factories have accumulated 20kg of the pollutant, the gap is given by equation (1). Note that under both profiles in this set, factory A is second-order responsible for the death of the fish because it has had an action ("do not dump") that would guarantee that the resulting profile does not belong to set G. By not dumping the pollutant, factory A guarantees that someone (factory B in our case) will be counterfactu-

ally responsible if the fish dies. Intuitively, second-order responsibility is the responsibility for the fact that no one is accountable.

Second-Order Gap   Let us now consider a setting where there is a third factory, C. Each of the three factories has accumulated 20kg of pollutant and it still takes only 15 kg of the pollutant to kill the fish. Thus, the deontic constraint γ is $\neg(p_a \vee p_b \vee p_c)$. Suppose that the factories make their decisions in the order: A, B, C. If either factory A or B dumps the pollutant, then the fish dies and no agent has had a strategy that would guarantee that the fish stays alive. Hence, nobody is counterfactually responsible for the death. If only factory C dumps the pollutant, then it has had a strategy ("do not dump") that would guarantee that the fish stays alive. Hence, in such a situation, factory C is counterfactually responsible for the death. In other words, the responsibility gap is

$$G_1 = \{(0,1,p_c) \,|\, p_c \in \{0,1\}\} \cup \{(1,p_b,p_c) \,|\, p_b,p_c \in \{0,1\}\}.$$

If the profile belongs to the set $\{(0,1,p_c) \mid p_c \in \{0,1\}\}$, then factory B is second-order responsible because it has a strategy ("do not dump") that guarantees that the decision-making process will avoid gap G_1. This is because this strategy guarantees that if the fish is dead, then factory C is responsible for the death. At the same time, if the profile belongs to the set

$$G_2 = \{(1,p_b,p_c) \mid p_b,p_c \in \{0,1\}\}, \tag{2}$$

then no agent has a strategy to guarantee that gap G_1 is avoided. Thus, under each profile from set G_2, not only the fish is dead and nobody is counterfactually responsible for this, but nobody is second-order responsible for the death either. We say that set G_2 is the *second-order responsibility gap*. Informally, the second-order responsibility gap consists of all profiles under which the deontic constraint is violated and nobody is responsible for the fact that nobody is accountable for the violation.

As a less formal example, imagine a situation when a paper submitted to a conference has not been reviewed by the PC member it got assigned to. In this case, the PC member is responsible for the failure to review. However, if no PC member has been assigned to review the paper, then there is a responsibility gap and the PC Chair is second-order responsible for the failure to review the paper. If the PC Chair gets ill and nobody has been assigned to replace the Chair, then there is a second-order gap.

First Result   Note that, in our three-factory example, factory A has a strategy ("do not dump") that guarantees that gap G_2 is avoided, see equation (2). Thus, under any profile that belongs to set G_2, factory A is *third-order responsible* for the death of the fish. One can also define a third-order gap as a set of all profiles in the second-order gap under which nobody is third-order responsible for a violation of the deontic constraint. In our example, the third-order gap is empty. In this paper, we consider n-order responsibility and n-order gaps. It is not a coincidence that, in our example with three agents, the third-order gap is empty. In Theorem 2, we show that in consecutive decision-making mechanisms with n agents, the n-order gap is always empty. In

other words, **if a deontic constraint is violated in an n-agent mechanism, then there exists $d \leq n$ such that at least one agent is d-order responsible for the violation**. The notion of n-order responsibility gap in a more general "extensive form game" setting is introduced in (Shi 2024). Shi showed that in that more general setting, the d-order gap is empty if $d \geq N - 1$, where N is the number of leaf nodes in the extensive form game. When translated to our setting, this means that the gap of order $2^n - 1$ is always empty. The reason why our result is so much stronger is that, in our setting, each agent acts just once. In an extensive form game, the same agent might act multiple times. Shi (2024) is also using the term "higher-order responsibility" in a sense similar but more general than ours.

Second Result Let us now consider a setting where it still takes 15kg of the pollutant to kill the fish, but factories A, B, and C have accumulated 20kg, 10kg, and 10kg of the pollutant, respectively. The deontic constraint is now $\gamma = \neg(p_a \vee (p_b \wedge p_c))$. In such a setting, the fish dies under the following set of action profiles:

$$\{(1, p_b, p_c) \mid p_b, p_c \in \{0, 1\}\} \cup \{(0, 1, 1)\}.$$

Out of these profiles, only under the last profile, somebody (actually both, factory B and factory C) is counterfactually responsible for the death of the fish. Thus, the responsibility gap is
$$G'_1 = \{(1, p_b, p_c) \mid p_b, p_c \in \{0, 1\}\}.$$

Note that factory A has a strategy ("do not dump") to avoid this gap. Thus, factory A is counterfactually responsible for the gap under each profile in set G'_1. This means that, in this setting, the second-order responsibility gap is empty. As we have seen in equation (2), in other settings such a gap might be nonempty. Thus, in some settings, the d-order responsibility gap might be empty for d strictly less than n. **Our second result is the complexity analysis of deciding for any given mechanism if the d-order responsibility gap is empty. In Theorem 3, we show that this problem is Π_{2d+1}-complete.** Shi (2024) discussed the algorithmic complexity of model-checking of an arbitrary formula that contains counterfactual responsibility modality. The absence of the d-order responsibility gap can be expressed by such a formula. Shi proved that model-checking can be done in polynomial time as a function of the number of leaf nodes in the tree representing the mechanism. In the case of the mechanisms discussed in the current paper, the number of leaf nodes is exponential in terms of the number of agents. Thus, when translated into our setting, her result only gives an exponential upper bound on the complexity of deciding if the d-order responsibility gap is empty.

Group Responsibility The main alternative to higher-order responsibility is *group responsibility* (Naumov and Tao 2019; List 2021; Yazdanpanah et al. 2021a,b; Dastani and Yazdanpanah 2023; Yazdanpanah et al. 2023). As an example, consider our three-factory scenario above: each factory has accumulated 20kg of pollutant and it takes 15kg to kill the fish. Suppose that the first factory (A) does *not* dump the pollutant and the other two factories dump (first B, then

C). The fish is killed. Note that neither of the three factories had an individual strategy that would guarantee saving the fish. Hence, none of them is counterfactually (first-order) responsible for killing the fish. At the same time, after factory A did not dump the pollutant, factories B and C had an opportunity to *coordinate* their actions and agree that neither of them would dump. This is a potential *group strategy* of B and C to guarantee that the fish is alive. Of course, the group consisting of all three factories also had a strategy to save the fish. The standard approach to define counterfactual group responsibility is to blame only the *minimal* (in terms of subset relation) group that could have prevented (Naumov and Tao 2020b). In our example, such a group consists of just B and C. Note that group responsibility diffuses the responsibility between B and C, potentially creating a "circle of blame" effect. Note that this approach is different from Necessary Element of a Sufficient Set causality by events or by specific actions (Wright 1985), rather than agents.

At the same time, in this example, only factory B is individually second-order responsible for the death of the fish. Indeed, if factory B refrains from dumping the pollutant, it will prevent the first-order responsibility gap (if C dumps, C will be counterfactually responsible). As we see from this example, higher-order responsibility avoids the diffusion associated with group responsibility and promotes individual accountability for the outcome.

"Responsibility Gap" in the Literature The responsibility gap is also often referred to as "responsibility void". These terms are consistently used in the literature to refer to a lack of a responsible agent in many different contexts. An important issue in AI ethics is whether autonomous systems should be treated as having "moral agency". If such systems are denied the moral agency status, then there is a commonly acknowledged responsibility gap for the decisions taken by the 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). The same terms are also used in group responsibility discussions when the group is responsible for an outcome but each individual agent is not (List 2021; Yazdanpanah et al. 2021a,b; Dastani and Yazdanpanah 2023; Yazdanpanah et al. 2023). Finally, outside moral agency and group responsibility contexts, these terms are used in decision-making mechanism analysis to refer to the outcomes that no agent is individually responsible for (Braham and van Hees 2011; Duijf 2018; Braham and van Hees 2018; Duijf and van De Putte 2022; Duijf 2022). Our work falls into the third category.

The responsibility gap has been studied in detail for "discursive dilemma" (List 2006) – a group decision-making mechanism where multiple agents are asked to independently rank available alternatives (say, job candidates) based on several required criteria. There might be a situation where each agent finds the alternative unacceptable due to one of the criteria, but the majority finds that the alternative satisfies all of the criteria. Ascribing individual responsibility for the group choice of the alternative in such a situation is problematic. (Braham and van Hees 2018; Duijf and van De Putte

2022) studied responsibility gaps in "discursive dilemma"-like settings. In such a setting, it is almost impossible for an agent to have a strategy that *guarantees* that an alternative is avoided. So, in both of these works Frankfurt's ". . . could have done otherwise" is interpreted as not taking an action which is the *most effective* in preventing the outcome. Braham and van Hees show that if a "discursive dilemma"-like setting has no responsibility gap and does not have what the authors call "fragmentation" of responsibility, then the mechanism must be a dictatorship, where a single agent determines the group decision (Braham and van Hees 2018). (Duijf and van De Putte 2022) gives necessary and sufficient conditions for a responsibility gap to exist in a "discursive dilemma"-like setting.

Formal Setting

In this section, we describe the formal setting in which we state and prove our technical results. We assume some fixed infinite countable set of propositional variables. By a Boolean formula, we mean any expression built out of propositional variables using conjunction $\wedge$, disjunction $\vee$, and negation $\neg$. By a Boolean formula with quantifiers, we mean a formula that, in addition, can use universal $\forall$ and existential $\exists$ quantifiers over propositional variables. If $\mathbf{v}$ is an ordered set of propositional variables $v_1, \ldots, v_k$, then by $\forall \mathbf{v}\varphi$ and $\exists \mathbf{v}\varphi$ we mean the formulae $\forall v_1 \ldots \forall v_k \varphi$ and $\exists v_1 \ldots \exists v_k \varphi$, respectively.

Definition 1 *A (sequential decision-making) mechanism is a triple $(n, \mathbf{v}, \gamma)$, where*

1. *integer $n \geq 0$ is the number of "agents",*
2. *$\mathbf{v} = \{\mathbf{v}_i\}_{1 \leq i \leq n}$ is a family of disjoint ordered sets of propositional variables; by $\cup \mathbf{v}$ we denote the unordered union of these sets,*
3. *"deontic constraint" γ is a Boolean formula (without quantifiers) whose variables belong to the set $\cup \mathbf{v}$.*

In our introductory examples, agents were factories A, B, and C. In the formal setting of Definition 1, agents are numbers 1, . . . , n. In those examples, we have assumed that the action of each agent i is represented by a single propositional variable p_i. In the more general setting of Definition 1, an action of each agent i is represented by an ordered set $\mathbf{v}_i$ of Boolean variables. To capture our introductory examples, we assume that $\mathbf{v}_1 = \{p_a\}$, $\mathbf{v}_2 = \{p_b\}$, and $\mathbf{v}_3 = \{p_c\}$.

Next, we will express the d-order responsibility of an agent i for the violation of the deontic constraint as a Boolean formula. Let us start with our introductory example where each of the three factories has accumulated 20kg of the pollutant. In this situation, the deontic constraint is $\gamma = \neg(p_a \vee p_b \vee p_c)$. Because factory A makes the decision first, for it to be (counterfactually) responsible, (i) the constraint should be violated and (ii) factory A should have an action that guarantees the constraint no matter what the actions of the other agents are. These two conditions can be expressed as

1. $p_a \vee p_b \vee p_c$, where p_a, p_b, p_c are unbounded variables representing the actual actions,

2. $\exists p_a \forall p_b \forall p_c \neg(p_a \vee p_b \vee p_c)$, where p_a, p_b, and p_c are bounded variables ranging over all possible actions.

Thus, the responsibility of factory A under action profile (p_a, p_b, p_c) can be expressed by the formula:

$$\mathsf{R}_a = (p_a \vee p_b \vee p_c) \wedge \exists p_a \forall p_b \forall p_c \neg(p_a \vee p_b \vee p_c). \quad (3)$$

Note that it is common in mathematics to avoid using the same variable names for bounded and unbounded variables. To do this, one can re-write formula R_a in an equivalent form. For example, as $(p_a \vee p_b \vee p_c) \wedge \exists x \forall y \forall z \neg(x \vee y \vee z)$. In this paper, we decided to define formula R_a as given by equation (3) because it makes it significantly simpler to express d-order responsibility later. It is important to remember, however, that there is no connection between identically named variables inside and outside of the scope of a corresponding quantifier.

Formula R_a expresses (counterfactual) responsibility of factory A in the sense that, for any action profile, if the values of unbounded variables p_a, p_b, and p_c are chosen as specified by the action profile, then the above formula has value 1 if and only if factory A is responsible for the violation of the deontic constraint γ.

Recall, that factory B makes its decision after factory A. Thus, for factory B to be (counterfactually) responsible, (i) the deontic constraint should be violated and (ii) factory B should have an action that guarantees the constraint **under the current action of factory A** and any possible action of factory C. This can be expressed as:

$$\mathsf{R}_b = (p_a \vee p_b \vee p_c) \wedge \exists p_b \forall p_c \neg(p_a \vee p_b \vee p_c). \quad (4)$$

Note that both occurrences of variable p_a in the above formula are unbounded. Thus, they both refer to the actual action of factory A. Similarly, the counterfactual responsibility of factory C is expressed by the formula

$$\mathsf{R}_c = (p_a \vee p_b \vee p_c) \wedge \exists p_c \neg(p_a \vee p_b \vee p_c). \quad (5)$$

More generally, under an arbitrary decision-making mechanism $(n, \mathbf{v}, \gamma)$, the (counterfactual) responsibility of agent $i \leq n$ is expressible by the formula:

$$\mathsf{R}_i = \neg\gamma \wedge \exists \mathbf{v}_i \forall \mathbf{v}_{i+1} \ldots \forall \mathbf{v_n} \gamma. \quad (6)$$

Then, the responsibility gap can be expressed by the formula

$$\mathsf{G} = \neg\gamma \wedge \bigwedge_{j \leq n} \neg\mathsf{R}_j.$$

The above formula expressed the gap in the sense that if unbounded occurrences of variables from ordered sets $\mathbf{v}_1, \ldots, \mathbf{v}_n$ represent the actions of the agents under some profile, then this formula has Boolean value 1 if and only if the profile belongs to the gap. Recall that the second-order responsibility for the violation of the deontic constraint is the counterfactual responsibility for the gap. Thus, it can be expressed by the formula

$$\mathsf{R}_i^2 = \left(\neg\gamma \wedge \bigwedge_{j \leq n} \neg\mathsf{R}_j \right)$$

$$\wedge \exists \mathbf{v}_i \forall \mathbf{v}_{i+1} \ldots \forall \mathbf{v}_n \neg \left(\neg\gamma \wedge \bigwedge_{j \leq n} \neg\mathsf{R}_j \right).$$

The third-order responsibility is expressible by the formula

$$
\mathsf{R}_i^3 = \left(\neg\gamma \wedge \bigwedge_{j \leq n} \neg\mathsf{R}_j \wedge \bigwedge_{j \leq n} \neg\mathsf{R}_j^2 \right) \wedge
$$

$$
\exists\mathbf{v}_i \forall\mathbf{v}_{i+1} \ldots \forall\mathbf{v}_n \neg \left(\neg\gamma \wedge \bigwedge_{j \leq n} \neg\mathsf{R}_j \wedge \bigwedge_{j \leq n} \neg\mathsf{R}_j^2 \right).
$$

In general, for any order $d > 0$, responsibility of agent i of order d is expressible by the formula

$$
\mathsf{R}_i^d = \left(\neg\gamma \wedge \bigwedge_{j \leq n} \neg\mathsf{R}_j \wedge \bigwedge_{j \leq n} \neg\mathsf{R}_j^2 \wedge \cdots \wedge \bigwedge_{j \leq n} \neg\mathsf{R}_j^{d-1} \right) \wedge
$$

$$
\exists\mathbf{v}_i \forall\mathbf{v}_{i+1} \ldots \forall\mathbf{v}_n
$$

$$
\neg \left(\neg\gamma \wedge \bigwedge_{j \leq n} \neg\mathsf{R}_j \wedge \bigwedge_{j \leq n} \neg\mathsf{R}_j^2 \wedge \cdots \wedge \bigwedge_{j \leq n} \neg\mathsf{R}_j^{d-1} \right). \quad (7)
$$

Recall that $\mathbf{v} = \{\mathbf{v}_i\}_{i \leq n}$ is a family of disjoint ordered sets of propositional variables. By an *action profile* we mean any family $\mathbf{s} = \{\mathbf{s}_i\}_{i \leq n}$ of ordered sets of Boolean values (0 or 1) such that $|\mathbf{s}_i| = |\mathbf{v}_i|$ for each agent $i \leq n$. For any formula φ whose unbounded propositional variables belong to the set $\cup \mathbf{v}$, let $\varphi[\mathbf{s}/\mathbf{v}]$ be the Boolean value (0 or 1) of formula φ under the valuation that assigns each variable from set $\mathbf{v}_i$ the corresponding Boolean value from set $\mathbf{s}_i$.

Definition 2 *Under an action profile $\mathbf{s}$ of a mechanism $(n, \mathbf{v}, \gamma)$, an agent $i \leq n$ is d-order responsible (for the violation of the deontic constraint) if $\mathsf{R}_i^d[\mathbf{s}/\mathbf{v}] = 1$.*

Definition 3 *A mechanism is d-gap-free if, for each action profile that does not satisfy the deontic constraint, there is $d' \leq d$ and an agent $i \leq n$ who is d'-order responsible under the action profile.*

The set of all d-gap-free mechanisms is denoted by GF^d.

Theorem 1 $\mathsf{GF}^{d_1} \subseteq \mathsf{GF}^{d_2}$ *for each $d_1 \leq d_2$.*

PROOF. By Definition 3 and the assumption $d_1 \leq d_2$, any d_1-gap-free mechanism is a d_2-gap-free mechanism. Therefore, $\mathsf{GF}^{d_1} \subseteq \mathsf{GF}^{d_2}$. $\qquad\square$

First Result

In this section, we prove our first result that any non-trivial mechanism with n agents is n-gap-free. This result is stated as Theorem 2 at the end of this section. We start with a lemma, whose proof is in the full version of the paper.

Lemma 1 *For any mechanism $(n, \mathbf{v}, \gamma)$, any nonnegative $k \leq n$, any action profile $\mathbf{s}$ such that $\gamma[\mathbf{s}/\mathbf{v}] = 1$, and any action profile $\mathbf{s}'$ such that $\mathbf{s}_i = \mathbf{s}'_i$ for each $i \leq k$, either $\gamma[\mathbf{s}'/\mathbf{v}] = 1$ or there is positive $d \leq n - k$ and an agent $i \leq n$ who is d-order responsible under action profile $\mathbf{s}'$.*

Theorem 2 *If Boolean formula γ is satisfiable, then $(n, \mathbf{v}, \gamma) \in \mathsf{GF}^n$.*

PROOF. The assumption that the Boolean formula γ is satisfiable implies that there is an action profile $\mathbf{s}$ such that $\gamma[\mathbf{s}/\mathbf{v}] = 1$. Consider an action profile $\mathbf{s}'$ such that $\gamma[\mathbf{s}'/\mathbf{v}] = 0$. By Definition 3, it suffices to show that there is $d \leq n$ and an agent $i \leq n$ who is d-order responsible under action profile $\mathbf{s}'$. The last statement follows from Lemma 1 in the case $k = 0$. $\qquad\square$

Together with Theorem 1, the above result implies that if γ is satisfiable, then $(n, \mathbf{v}, \gamma) \in \mathsf{GF}^d$ for each $d \geq n$.

Second Result

In this section, we prove that set GF^d is Π_{2d+1}-complete. The result is stated as Theorem 3 at the end of the section.

Lemma 2 *For each $d \geq 1$, set GF^d belongs to the class Π_{2d+1}.*

PROOF. First, observe that the following claim holds.

Claim 1 *For each $d \geq 1$ and each $i \leq n$, formula R_i^d belongs to class Σ_{2d}.*

PROOF OF CLAIM. We prove the claim by induction on d. If $d = 1$, then R_i^d is the formula $\neg\gamma \wedge \exists\mathbf{v}_i \forall\mathbf{v}_{i+1} \ldots \forall\mathbf{v_n}\gamma$ by equation (7) (or equation (6)). Note that γ is a Boolean formula without quantifiers by item 3 of Definition 1. Hence, formula R_i^d belongs to class Σ_2. The induction step follows from equation (7) and the induction hypothesis. $\qquad\boxtimes$

By Definition 3 and Definition 2, a mechanism $(n, \mathbf{v}, \gamma)$ is d-gap-free if and only if the following closed formula is a tautology:

$$
\forall\mathbf{v}_1 \ldots \forall\mathbf{v}_n \left(\gamma \vee \bigvee_{d' \leq d} \bigvee_{j \leq n} \mathsf{R}_j^{d'} \right).
$$

By Claim 1, this formula belongs to class Π_{2d+1}, which implies the statement of the lemma. $\qquad\square$

Let us now prove that the set GF^d is Π_{2d+1}-complete by defining a polynomial reduction of Π_{2d+1} problems to GF^d. Towards this goal, for each Boolean formula, we specify a decision-making mechanism.

Definition 4 *For any $d \geq 0$, any Boolean formula φ without quantifiers, and any family of disjoint ordered sets $\mathbf{x}_1, \ldots, \mathbf{x}_{2d+1}$ of propositional variables whose union is equal to the set of all propositional variables in formula φ, mechanism $M(\varphi, \mathbf{x}_1, \ldots, \mathbf{x}_{2d+1}) = (2d+1, \mathbf{v}, \gamma)$ is defined as follows:*

1. *$\mathbf{v}_{2i+1} = \mathbf{x}_{2i+1}$ for each i such that $0 \leq i \leq d$,*
2. *$\mathbf{v}_{2i}$ is equal to the concatenation of a new single variable q_{2i} to the end of the ordered list of variables $\mathbf{x}_{2i}$, for each i such that $1 \leq i \leq d$,*
3. *$\gamma = \varphi \wedge \bigwedge_{1 \leq i \leq d} q_{2i}$.*

As an example, suppose that $d = 1$, formula φ is the formula $x_1 \vee x_2 \vee x_3$, $\mathbf{x}_1 = \{x_1\}$, $\mathbf{x}_2 = \{x_2\}$, and $\mathbf{x}_3 = \{x_3\}$. Then, mechanism $M(\varphi, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ has three agents (called agent 1, agent 2, and agent 3). The action of agent 1 is defined by the singleton set of variables $\mathbf{v}_1 = \mathbf{x}_1 = \{x_1\}$. The action of

agent 2 is defined by the ordered set $\mathbf{v}_2 = \{x_2, q_2\}$, where q_2 is a new propositional variable that does not appear among variables x_1, x_2, and x_3. The action of agent 3 is defined by $\mathbf{v}_3 = \mathbf{x}_3 = \{x_3\}$. The deontic constraint γ is $(x_1 \vee x_2 \vee x_3) \wedge q_2$.

The connection between formula φ and the decision-making mechanism $M(\varphi, \mathbf{x}_1, \ldots, \mathbf{x}_{2d+1})$ is stated in Lemma 7 and Lemma 8 below. To prove both of these lemmas, we introduce a *two-player extensive form game* between Devil and Moralist. In the game, Devil is in charge of odd-numbered sets $\mathbf{v}_1, \mathbf{v}_3, \mathbf{v}_5, \ldots, \mathbf{v}_{2d+1}$ and Moralist is in charge of even-numbered sets $\mathbf{v}_2, \mathbf{v}_4, \ldots, \mathbf{v}_{2d}$. The game consists of Devil and Moralist consequently choosing the values of variables in ordered sets they control: first, Devil chooses the values of all variables in $\mathbf{v}_1$, then Moralist chooses the values of all variables in $\mathbf{v}_2$, then Devil chooses $\mathbf{v}_3$, then Moralist chooses $\mathbf{v}_4$, and so on. Devil always concludes the game by choosing the values of all variables in set $\mathbf{v}_{2d+1}$. Once the game is over, they have specified an action profile of the mechanism $M(\varphi, \mathbf{x}_1, \ldots, \mathbf{x}_{2d+1})$. The Devil's objective is to violate the deontic constraint γ on the action profile they specified. The objective of Moralist is to satisfy γ.

Figure 1 shows the game between Devil and Moralist for the example we discussed above. The game starts at a root node ε with none of the variables specified. Then, Devil can choose the value of x_1 to be either 0 (go left) or 1 (go right). Then, Moralist chooses the values of x_2 and q_2. This corresponds to a transition from node x_1 to node $x_1 x_2 q_2$ in the tree. Finally, Devil picks the value of x_3, which completely specifies the action profile $x_1 x_2 q_2 x_3$ of the decision-making mechanism. Note that, to keep the drawing clean, in Figure 1 we write $x_1 x_2 q_2 x_3$ instead of more accurate $(x_1, (x_2, q_2), x_3)$. In the figure, we labelled with γ the leaf nodes whose action profiles satisfy the formula $\gamma = (x_1 \vee x_2 \vee x_3) \wedge q_2$. These are the outcomes of the game in which Moralist prevails.

It is easy to see that in our example in Figure 1, Moralist has a *strategy* to guarantee γ. By Win(Moralist) we denote the set of all (not necessarily leaf) nodes in which Moralist has a *strategy* to guarantee the deontic constraint. For the example depicted in Figure 1,

$$\text{Win(Moralist)} = \{\varepsilon, 0, 1, 011, 101, 111, 0011,$$
$$0110, 0111, 1010, 1011, 1110, 1111\}.$$

In Lemma 6, we establish a connection between the responsibility of agents under the profiles of mechanism $M(\varphi, \mathbf{x}_1, \ldots, \mathbf{x}_{2d+1})$ and existence of winning strategies in the two-player game between Devil and Moralist. In preparation for proving Lemma 6, below we state three small observations about winning strategies in the game.

Lemma 3 *If* $(\mathbf{s}_1, \ldots, \mathbf{s}_{2k}) \in$ Win(Moralist)*, then* $(\mathbf{s}_1, \ldots, \mathbf{s}_{2k}, \mathbf{s}_{2k+1}) \in$ Win(Moralist)*.*

PROOF. According to the rules of the two-player game, it is Devil's turn to make a move at the node $(\mathbf{s}_1, \ldots, \mathbf{s}_{2k})$. Thus, in order for Moralist to have a winning strategy at this node, Moralist must have such a strategy after each possible

move of Devil. In particular, Moralist must have a winning strategy at the node $(\mathbf{s}_1, \ldots, \mathbf{s}_{2k}, \mathbf{s}_{2k+1})$. $\square$

The proofs of the next three lemmas can be found in the full version of the paper.

Lemma 4 *If* $(\mathbf{s}_1, \ldots, \mathbf{s}_{2k+1}) \in$ Win(Moralist)*, then there exists an action* $\mathbf{s}'_{2k+2}$ *of agent* $2k + 2$ *such that* $(\mathbf{s}_1, \ldots, \mathbf{s}_{2k+1}, \mathbf{s}'_{2k+2}) \in$ Win(Moralist)*.*

Lemma 5 *If* $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_{2d+1}) \in$ Win(Moralist)*, then* $\gamma[\mathbf{s}/\mathbf{v}] = 1$*.*

Lemma 6 *For any integer* k *such that* $0 \leq k \leq d$ *and any action profile* $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_{2d+1})$*, if* $(\mathbf{s}_1, \ldots, \mathbf{s}_{2k+1}) \in$ Win(Moralist)*, then either* $\gamma[\mathbf{s}/\mathbf{v}] = 1$ *or there is positive* $d' \leq d - k$ *and an agent* $i \leq 2d + 1$ *who is* d'*-order responsible under action profile* $\mathbf{s}$*.*

Lemma 7 *If the closed Boolean formula with quantifiers* $\forall \mathbf{x}_1 \exists \mathbf{x}_2 \forall \mathbf{x}_3 \ldots \forall \mathbf{x}_{2d+1} \varphi$ *is true, then, for any action profile* $\mathbf{s}$*, either* $\gamma[\mathbf{s}/\mathbf{v}] = 1$ *or there is positive* $d' \leq d$ *and an agent* $i \leq 2d + 1$ *who is* d'*-order responsible under profile* $\mathbf{s}$*.*

PROOF. Suppose that the closed formula $\forall \mathbf{x}_1 \exists \mathbf{x}_2 \forall \mathbf{x}_3 \ldots \forall \mathbf{x}_{2d+1} \varphi$ is true. Thus, the formula

$$\forall \mathbf{x}_1 \exists \mathbf{x}_2 \exists q_2 \forall \mathbf{x}_3 \exists \mathbf{x}_4 \exists q_4 \ldots \exists \mathbf{x}_{2d} \exists q_{2d} \forall \mathbf{x}_{2d+1}$$
$$(\varphi \wedge q_2 \wedge q_4 \wedge \cdots \wedge q_{2d})$$

is also true because the values of $q_2, q_4, \ldots, q_{2d}$ could be chosen to be 1. Then, by Definition 4, the formula $\forall \mathbf{v}_1 \exists \mathbf{v}_2 \forall \mathbf{v}_3 \ldots \forall \mathbf{v}_{2d+1} \gamma$ is true. Recall that Moralist is in charge of even-numbered sets $\mathbf{v}_2, \mathbf{v}_4, \ldots, \mathbf{v}_{2d}$. Thus, at the initial moment in the game, Moralist has a strategy to guarantee that the deontic constraint γ is satisfied. In other words, $\varepsilon \in$ Win(Moralist). Suppose $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_{2d+1})$. Then, $(\mathbf{s}_1) \in$ Win(Moralist) by Lemma 3. Therefore, by Lemma 6 for $k = 0$, either $\gamma[\mathbf{s}/\mathbf{v}] = 1$ or there is $d' \leq d$ and an agent $i \leq 2d + 1$ who is d'-order responsible under action profile $\mathbf{s}$. $\square$

Lemma 8 *If the closed Boolean formula with quantifiers* $\forall \mathbf{x}_1 \exists \mathbf{x}_2 \forall \mathbf{x}_3 \ldots \forall \mathbf{x}_{2d+1} \varphi$ *is false, then there exists an action profile* $\mathbf{s}$ *such that* $\gamma[\mathbf{s}/\mathbf{v}] = 0$ *and there is no* $d' \leq d$ *such that at least one of the agents is* d'*-order responsible under profile* $\mathbf{s}$*.*

PROOF. The assumption that the closed formula $\forall \mathbf{x}_1 \exists \mathbf{x}_2 \forall \mathbf{x}_3 \ldots \forall \mathbf{x}_{2d+1} \varphi$ is false implies that the formula $\exists \mathbf{x}_1 \forall \mathbf{x}_2 \exists \mathbf{x}_3 \ldots \exists \mathbf{x}_{2d+1} \neg \varphi$ is true. Thus, the formula

$$\exists \mathbf{x}_1 \forall \mathbf{x}_2 \forall q_2 \exists \mathbf{x}_3 \forall \mathbf{x}_4 \forall q_4 \ldots \forall \mathbf{x}_{2d} \forall q_{2d} \exists \mathbf{x}_{2d+1}$$
$$\neg(\varphi \wedge q_2 \wedge q_4 \wedge \cdots \wedge q_{2d})$$

is also true because to make a conjunction false it suffices to guarantee that the first conjunct is false. Then, by Definition 4, the formula $\exists \mathbf{v}_1 \forall \mathbf{v}_2 \exists \mathbf{v}_3 \ldots \exists \mathbf{v}_{2d+1} \neg \gamma$ is true. Recall that Devil is in charge of odd-numbered sets $\mathbf{v}_1, \mathbf{v}_3, \ldots, \mathbf{v}_{2d+1}$. Thus, at the initial moment in the game, Devil has (at least one) strategy to guarantee that the deontic constraint γ is violated. Let σ be one of such strategies.

For any action profile $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_{2d+1})$, we define the "degree of immorality" $\|\mathbf{s}\|$ of profile $\mathbf{s}$ as the total number of "sins" committed under profile $\mathbf{s}$. The "sins" are defined as follows:
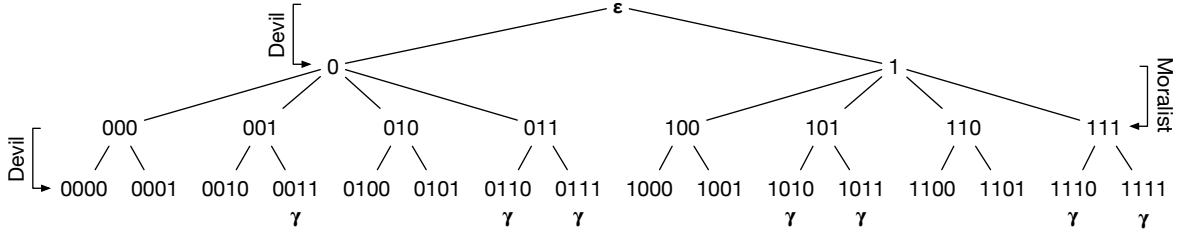
Figure 1: Extensive form game between Devil and Moralist.

1. each $i \leq 2d+1$ such that $q_{2i}[\mathbf{s}/\mathbf{v}] = 0$ is a "sin",
2. it is a "sin" if Devil follows strategy σ under the profile $\mathbf{s}_1, \ldots, \mathbf{s}_{2d+1}$.

Note that $\gamma = \varphi \wedge \bigwedge_{1 \leq i \leq d} q_{2i}$ by Definition 4. Thus, committing even a single "sin" under profile $\mathbf{s}$ guarantees that $\gamma[\mathbf{s}/\mathbf{v}] = 0$. Committing two sins "double-proofs" it. Committing three sins "triple-proofs" it. Intuitively, the "degree of immorality" $\|\mathbf{s}\|$ is the measurement of how "evil" is the behaviour of the agents $1, \ldots, 2d+1$ under a profile $\mathbf{s}$.

Claim 2 *For any profile $\mathbf{s}$, if $\gamma[\mathbf{s}/\mathbf{v}] = 1$, then $\|\mathbf{s}\| = 0$.*

PROOF OF CLAIM. The statement of the claim holds because committing even a single "sin" guarantees that $\gamma[\mathbf{s}/\mathbf{v}] = 0$. ⊠

Claim 3 *For any action profiles $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_{2d+1})$, any $i \leq 2d+1$ and any $\mathbf{s}'_i$, there is an action profile $\mathbf{s}' = (\mathbf{s}_1, \ldots, \mathbf{s}_{i-1}, \mathbf{s}'_i, \mathbf{s}_{i+1}, \ldots, \mathbf{s}_{2d+1})$ such that $\|\mathbf{s}\| - 1 \leq \|\mathbf{s}'\|$.*

PROOF OF CLAIM. If number i is odd, then let $\mathbf{s}'$ be the action profile $(\mathbf{s}_1, \ldots, \mathbf{s}_{i-1}, \mathbf{s}'_i, \mathbf{s}_{i+1}, \ldots, \mathbf{s}_{2d+1})$. Note that $q_{2j}[\mathbf{s}/\mathbf{v}] = q_{2j}[\mathbf{s}'/\mathbf{v}]$ for each $j \leq d$ because i is odd. Therefore, $\|\mathbf{s}\| - 1 \leq \|\mathbf{s}'\|$ by the definition of the degree of immorality.

If i is even, then Devil is *not* in charge of set $\mathbf{v}_i$. Thus, there is an action profile

$(\mathbf{s}_1, .., \mathbf{s}_{i-1}, \mathbf{s}'_i, \mathbf{s}'_{i+1}, \mathbf{s}_{i+2}, \mathbf{s}'_{i+3}, \mathbf{s}_{i+4}, \mathbf{s}'_{i+5}, .., \mathbf{s}_{2d}, \mathbf{s}'_{2d+1})$

under which Devil still follows strategy σ if Devil followed σ under profile $\|\mathbf{s}\|$. Let $\mathbf{s}'$ be such a profile. Then, $\|\mathbf{s}\| - 1 \leq \|\mathbf{s}'\|$ because $q_{2j}[\mathbf{s}/\mathbf{v}] = q_{2j}[\mathbf{s}'/\mathbf{v}]$ when $2j \neq i$. ⊠

Claim 4 *For any profile $\mathbf{s}$, any positive integer $k < \|\mathbf{s}\|$, and any agent $i \leq 2d+1$, agent i is not k-order responsible under action profile $\mathbf{s}$.*

PROOF OF CLAIM. We prove the statement by induction on $\|\mathbf{s}\|$. If $\|\mathbf{s}\| \leq 1$, then the statement vacuously holds because there are no positive integers k such that $k < 1$.

Suppose that there is an integer k such that

$$1 < k < \|\mathbf{s}\| \qquad (8)$$

and a k-order responsible agent $i \leq 2d+1$ under $\mathbf{s}$. Thus, $\mathsf{R}_i^k[\mathbf{s}/\mathbf{v}] = 1$ by Definition 2. Hence, by equation (7),

$\exists \mathbf{v}_i \forall \mathbf{v}_{i+1} \ldots \forall \mathbf{v}_{2d+1}$

$$\neg \left(\neg \gamma \wedge \bigwedge_{j \leq 2d+1} \neg \mathsf{R}_j \wedge \cdots \wedge \bigwedge_{j \leq 2d+1} \neg \mathsf{R}_j^{k-1} \right) [\mathbf{s}/\mathbf{v}] = 1.$$

Let $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_{2d+1})$. Then, there is $\mathbf{s}'_i$ such that

$$\neg \left(\neg \gamma \wedge \bigwedge_{j \leq 2d+1} \neg \mathsf{R}_j \cdots \wedge \bigwedge_{j \leq 2d+1} \neg \mathsf{R}_j^{k-1} \right) [\mathbf{s}'/\mathbf{v}] = 1$$

for *each* profile $\mathbf{s}' = (\mathbf{s}_1, \ldots, \mathbf{s}_{i-1}, \mathbf{s}'_i, \mathbf{s}'_{i+1}, \ldots, \mathbf{s}'_{2d+1})$. By Claim 3, profile $\mathbf{s}'$ can be *chosen* in such a way that

$$\|\mathbf{s}\| - 1 \leq \|\mathbf{s}'\|. \qquad (9)$$

Then,

$$\left(\gamma \vee \bigvee_{j \leq 2d+1} \mathsf{R}_j \vee \cdots \vee \bigvee_{j \leq 2d+1} \mathsf{R}_j^{k-1} \right) [\mathbf{s}'/\mathbf{v}] = 1. \quad (10)$$

Inequalities (8) and (9) imply that $\|\mathbf{s}'\| \geq k > 1$. Thus, $\gamma[\mathbf{s}'/\mathbf{v}] = 0$ by Claim 2. Hence, by equation (10), there is $j \leq 2d+1$ and $k' \leq k-1$ such that $\mathsf{R}_j^{k'}[\mathbf{s}'/\mathbf{v}] = 1$. Then, by the induction hypothesis, $\|\mathbf{s}'\| < k'$. Thus, by inequality (9),

$$\|\mathbf{s}\| \leq \|\mathbf{s}'\| + 1 < k' + 1 \leq (k - 1) + 1 = k,$$

which contradicts inequality (8). ⊠

Consider any action profile $\mathbf{s}$ under which Devil follows strategy σ and $q_{2i}[\mathbf{s}/\mathbf{v}] = 0$ for each $i \leq d$. In other words, we consider an action profile under which all $d+1$ possible "sin"s have been committed. Thus,

$$\|\mathbf{s}\| = d + 1. \qquad (11)$$

Note that $\gamma[\mathbf{s}/\mathbf{v}] = 0$, because Devil's strategy σ guarantees that the deontic constraint is violated. To finish the proof of the lemma, suppose that there is $d' \leq d$ such that at least one agent is d'-order responsible under $\mathbf{s}$. Thus, $d' \geq \|\mathbf{s}\|$ by Claim 4. Therefore, $d \geq d' \geq \|\mathbf{s}\| = d+1$ by statement (11), which is a contradiction. □

The next result follows from the two previous lemmas.

Lemma 9 *Set GF^d is Π_{2d+1}-hard.*

Theorem 3 *Set GF^d is Π_{2d+1}-complete.*

PROOF. The statement of the theorem follows from Lemma 2 and Lemma 9. □

Conclusion

In this paper, we have studied higher-order responsibility gaps in sequential decision-making mechanisms. We have shown that the set of d-order gap-free mechanisms is Π_{2d+1}-complete. At the same time, the mechanism is guaranteed to be d-order gap-free when the number of agents is at most d.

Acknowledgments

References

Baier, C.; Funke, F.; and Majumdar, R. 2021. A Game-Theoretic Account of Responsibility Allocation. In *30th International Joint Conference on Artificial Intelligence (IJCAI-21)*.

Blumenfeld, D. 1971. The principle of alternate possibilities. *The Journal of Philosophy*, 68(11): 339–345.

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

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

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

Champagne, M.; and Tonkens, R. 2015. Bridging the responsibility gap in automated warfare. *Philosophy & Technology*, 28: 125–137.

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

Copp, D. 1997. Defending the principle of alternate possibilities: Blameworthiness and moral responsibility. *Nous*, 31(4): 441–456.

Dastani, M.; and Yazdanpanah, V. 2023. Responsibility of AI systems. *AI & Society*, 38(2): 843–852.

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

Duijf, H. 2022. *The logic of responsibility voids*. Springer.

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

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

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

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

Hindriks, F.; and Veluwenkamp, H. 2023. The risks of autonomous machines: from responsibility gaps to control gaps. *Synthese*, 201(1): 21.

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

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

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

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

Mele, A. R.; and Robb, D. 2018. Rescuing Frankfurt-Style Cases. In *Agency And Responsiblity*, 241–254. Routledge.

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

Naumov, P.; and Tao, J. 2020a. Blameworthiness in Security Games. In *Proceedings of Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*.

Naumov, P.; and Tao, J. 2020b. An Epistemic Logic of Blameworthiness. *Artificial Intelligence*, 283. 103269.

Naylor, M. B. 1984. Frankfurt on the principle of alternate possibilities. *Philosophical Studies*, 46(2): 249–258.

Oimann, A.-K. 2023. The responsibility gap and LAWS: A critical mapping of the debate. *Philosophy & Technology*, 36(1): 3.

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

Shi, Q. 2024. Responsibility in Extensive Form Games. In *Proceedings of 38th AAAI Conference on Artificial Intelligence (AAAI-24)*.

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

Widerker, D.; and McKenna, M., eds. 2003. *Moral responsibility and alternative possibilities: Essays on the importance of alternative possibilities*. Burlington, VT: Ashgate.

Wright, R. W. 1985. Causation in tort law. *Calif. L. Rev.*, 73: 1735.

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

Yazdanpanah, V.; Gerding, E. H.; Stein, S.; Cîrstea, C.; m. c. schraefel; Norman, T. J.; and Jennings, N. R. 2021a. Different Forms of Responsibility in Multiagent Systems: Sociotechnical Characteristics and Requirements. *IEEE Internet Comput.*, 25(6): 15–22.

Yazdanpanah, V.; Gerding, E. H.; Stein, S.; Dastani, M.; Jonker, C. M.; and Norman, T. J. 2021b. Responsibility Research for Trustworthy Autonomous Systems. In Dignum, F.; Lomuscio, A.; Endriss, U.; and Nowé, A., eds., *AAMAS '21: 20th International Conference on Autonomous Agents and Multiagent Systems, Virtual Event, United Kingdom, May 3-7, 2021*, 57–62. ACM.

Yazdanpanah, V.; Gerding, E. H.; Stein, S.; Dastani, M.; Jonker, C. M.; Norman, T. J.; and Ramchurn, S. D. 2023. Reasoning about responsibility in autonomous systems: challenges and opportunities. *AI & SOCIETY*, 38(4): 1453–1464.