A Game-Theoretic Account of Responsibility Allocation

Christel Baier[1] , **Florian Funke**[1] , **Rupak Majumdar**[2]
[1]Technische Universität Dresden, Dresden, Germany
[2]Max Planck Institute for Software Systems, Kaiserslautern, Germany
{christel.baier, florian.funke}@tu-dresden.de, rupak@mpi-sws.org

Abstract

When designing or analyzing multi-agent systems, a fundamental problem is *responsibility ascription*: to specify which agents are responsible for the joint outcome of their behaviors and to which extent. We model strategic multi-agent interaction as an extensive form game of imperfect information and define notions of forward (prospective) and backward (retrospective) responsibility. Forward responsibility identifies the responsibility of a group of agents for an outcome along all possible plays, whereas backward responsibility identifies the responsibility along a given play. We further distinguish between strategic and causal backward responsibility, where the former captures the epistemic knowledge of players along a play, while the latter formalizes which players—possibly unknowingly—caused the outcome. A formal connection between forward and backward notions is established in the case of perfect recall. We further ascribe *quantitative* responsibility through cooperative game theory. We show through a number of examples that our approach encompasses several prior formal accounts of responsibility attribution.

1 Introduction

The notion of *responsibility* is fundamental in the study of multi-agent interaction. Allocation of responsibility is a means by which we regulate interactions in society, by declaring whether an action by a person in an interactive setting should be praised or blamed. In multi-agent interactions, ascertaining who is to be held responsible and by which degree can be difficult; thus, there is a need for formal frameworks for responsibility allocation.

We work in the framework of "folk ethics" conception of moral responsibility [Braham and van Hees, 2012; Braham and van Hees, 2018]. In this setting, the locus of responsibility resides with the individual rather than a collective and purely on her actions and their consequences rather than identities, attitudes, norms, or values. A person is ascribed responsibility for a given outcome if three conditions are met. First, the person has *agency*: they are able to plan and act intentionally and can distinguish the setting of the interaction and outcomes. The second is *causal relevance*: there is a causal link between the actions and the outcome—to be elaborated below. The third is the *possibility to act otherwise*: We use our evaluation of actual and potential actions and its consequences on the outcome. However, our approach is descriptive in the sense that we do not analyze normative aspects such as values, virtues, intent, or morality underlying the actions of the agents [Scanlon, 1998].

In this paper, we provide a game-theoretic account of responsibility allocation in a multi-agent interaction setting. We model multi-agent interaction as a game of imperfect information in extensive form [Kuhn, 1953; Owen, 1995] between n individually rational players. Players are assumed to be rational in a weak sense: they are aware of the game, the other players, their own actions, and the outcome. In particular, we shall assume agency. We study both *forward* and *backward* notions of responsibility [van de Poel, 2011]. A forward notion looks at the game as a whole and ascribes responsibilities to players based on all potential plays. A backward notion looks at a play and ascribes responsibility to each player for that play. We approach responsibility allocation to individual agents in two steps. First, we look at coalitions of players and define notions of forward and backward responsibilities for a coalition. Second, we define a value function from coalitions to their responsibilities, and define the individual allocation as a power index of this value function [Owen, 1995].

For the first step, we distinguish between *causal* and *strategic* backward responsibility, which correspond to the *responsiblity-as-cause* and *responsibility-as-capacity* notions of [van de Poel, 2011]. Intuitively, a coalition is causally backward responsible for an outcome along a play if there is a different strategy the coalition could have adopted that would have avoided the outcome, against the same strategy of the other players. Strategic backward responsibility strengthens the requirement: the coalition should be aware, given their epistemic state, of this ability to affect the outcome. Our key technical result is a relationship between forward and strategic backward responsibility: in a game of perfect recall, a coalition is forward responsible for an outcome *iff* it contains a strategically backward responsible coalition for every play with the outcome, and is minimal with respect to this property (Theorem 1). Moreover, we show that all forms of responsibility of a coalition can be checked in polynomial time (Theorem 2).

For the second step, our approach follows [von Neumann, 1928], and relies on a transition from non-cooperative games to cooperative values. While there are different measures of value in a cooperative game, we pick the Shapley value [Shapley, 1953] for its familiarity and its canonicity. Thus, we ascribe *quantitative* measures of responsibility in order to compare the relative responsibility of agents for an outcome. As a motivation we show that our modeling choices allow us to precisely talk about various aspects of many well-known scenarios from the moral philosophy and causality literature.

The condition for causal dependence is often tested in a framework of *actual causality* [Halpern and Pearl, 2005; Halpern, 2015] (henceforth HP, after its proponents) or the NESS test [Hart and Honoré, 1959]. HP provides a formal model for causal contribution to an outcome and models institutional rules as *structural equations*. We believe that structural equations are too weak to naturally model several situations of interest, such as agency, knowledge, or temporal sequentiality. Consider the prototypical example in which Suzy and Billy both throw stones at a bottle. Suzy's stone hits first and the bottle breaks [Halpern, 2016]. Who should be responsible? The description of the problem or the structural equations do not clarify if either Suzy or Billy knew if the other threw a rock, or if Billy's strategy to throw was conditional, knowing that Suzy had indeed thrown already. These nuances are easily modeled in our setting due to the additional modeling power of extensive form games.

In a technical sense, we can embed structural equations into our framework, and our causal backward responsibility is exactly the *but-for* condition in causal reasoning (Theorem 3). HP's actual causality goes beyond but-for-causes: in the above example, HP considers Suzy to be responsible – but not Billy – by interposing a counterfactual world in which Suzy does *not* throw, Billy does throw, but still Billy's stone does *not* hit the bottle (where we use Halpern's modified version [Halpern, 2015]). We find this problematic: the intervention (Billy's stone not hitting the bottle) is not an agent to whom we can ascribe agency. Because of the symmetry of the players, and our insistence on comparing strategies with other agent strategies, we hold both equally responsible. We believe HP's allocation of backward responsibility solely to Suzy is not uncontroversial: replace Suzy and Billy with two assassins who simultaneously (and without knowledge of the other) shoot a person. Even if the laws of physics decide which bullet reaches first, in moral or legal considerations, we would hold both assassins responsible. A second advantage of games over structural equations is that responsibility of a coalition can be computed in polynomial time; thus, checking if an individual is responsible is in NP, as opposed to the harder class D^P for HP.

Full proofs can be found in the technical report [Baier *et al.*, 2021].

Other Related Work. Close to our work, [Braham and van Hees, 2012] give an account of moral responsibility as *normal form* games with pure strategies (the model is one-shot, perfect-information). Models of (qualitative and quantitative) responsibility have been studied for non-probabilistic Kripke structures with games of possibly infinite duration (see, e.g., [Chockler *et al.*, 2008; Beer *et al.*, 2009; Bulling and Dastani, 2013; Yazdanpanah and Dastani, 2016]). Quantitative measures of *influence* in causal models [Chockler and Halpern, 2004] have generated a fruitful strand in the causality literature [Aleksandrowicz *et al.*, 2014; Chockler, 2016; Friedenberg and Halpern, 2019]. A detailed comparison with other accounts of responsibility is the content of Section 5. Shapley-like values have been used to allocate responsibility [Friedenberg and Halpern, 2019; Yazdanpanah *et al.*, 2019], and recently been rediscovered for the explanation of machine learning models [Štrumbelj and Kononenko, 2014; Lundberg and Lee, 2017].

2 Preliminaries

In this section we recall some basic definitions from non-cooperative game theory (see, e.g., [Owen, 1995]). We rely on von Neumann's framework of extensive form games [von Neumann, 1928] with the only exception that leaves are not labeled with payoff vectors, but a binary variable E or $\neg E$ indicating whether a certain event has occured or not.

Definition 1 (Extensive form game)**.** *A finite* n-player game $\mathcal{G}$ *in* extensive form *consists of the following data:*

1. *a finite directed tree* $\mathcal{T}$*, called the* game tree*, whose vertices are called the* states *of* $\mathcal{G}$*;*
2. *a partition of the non-terminal states of* $\mathcal{T}$ *into sets* $P_0, P_1, ..., P_n$*;*
3. *for* $s \in P_i$ *a finite set of actions* $\mathrm{Act}(s)$ *that is in bijection with the successors of* s*;*
4. *for* $s \in P_0$ *a probability distribution* $p(s)$ *over* $\mathrm{Act}(s)$*;*
5. *a partition* $P_i = \dot{\bigcup}_{I \in \mathcal{I}_i} I$ *into* information sets *such that* $\mathrm{Act}(s) = \mathrm{Act}(s')$ *for all* s, s' *in the same information set* I*, and* $p(s) = p(s')$ *if additionally* $s, s' \in P_0$*;*
6. *for each leaf* s *of* $\mathcal{T}$ *a* labeling $l(s) \in \{E, \neg E\}$.

Intuitively, the game is played as follows. One starts at the root s_0 of $\mathcal{T}$. If the current state s belongs to P_0, then the next state is chosen by a random player called `Nature` according to $p(s)$. If the current state s belongs to P_i, then player i chooses a successor state by choosing an action from $\mathrm{Act}(s)$. The intended meaning of the information set I is that player i cannot distinguish the states in I and must choose an action independent of the specific state in I. A *play* in $\mathcal{G}$ is a path from the root to a terminal state, and the set of plays is denoted by $\mathrm{Plays}(\mathcal{G})$. The plays that run through state s or information set I are denoted by $\mathrm{Plays}_s(\mathcal{G})$ and $\mathrm{Plays}_I(\mathcal{G})$. A play ending in a state with label E is called an E*-play*.

Throughout, we will restrict our attention to games of *perfect recall*, in which, intuitively, no player forgets his own history of actions. This is a reasonable assumption in many real-world scenarios, especially those of short duration. It also captures the intuition in a setting of responsibility that an agent should have known information available. For a state s let $\mathrm{hist}_{\mathcal{G}}(s) = I_0 a_0 I_1 a_1 ... a_{k-1} I_k$ be the sequence of information sets visited and actions taken on the path from the root of $\mathcal{T}$ to s. Given $C \subseteq \{1, ..., n\}$ let $\mathrm{hist}_{\mathcal{G}}(s, C)$ be the subsequence obtained from $\mathrm{hist}_{\mathcal{G}}(s)$ by removing each I_j and action a_j that is not under control of a player in C.

Definition 2 (Perfect recall)**.** *A game $\mathcal{G}$ has* perfect recall *if for each player i and any two states s, s' in the same information set of player i we have* $\text{hist}_{\mathcal{G}}(s, \{i\}) = \text{hist}_{\mathcal{G}}(s', \{i\})$.

In the presence of uncertainty that comes with non-singleton information sets, players may prefer to act *randomly* instead of *deterministically*. We will be allowing behavioral strategies throughout [Kuhn, 1953].

Definition 3 (Strategies)**.** *A* (behavioral) strategy *for player i is an element $\sigma_i = \{\sigma_I\}_{I \in \mathcal{I}_i} \in \prod_{I \in \mathcal{I}_i} \text{Dist}(\text{Act}(I))$. It is* pure *if each σ_I is a Dirac distribution. A* strategy profile *is a set $\{\sigma_i\}_{i=1}^n$, where each σ_i is a strategy for player i.*

A play ρ is *consistent* with a strategy profile $\sigma = \{\sigma_i\}$ if for every information set I on ρ the chosen action has positive probability in σ_I. Taking only consistent plays induces subsets $\text{Plays}_s^{\sigma}(\mathcal{G}) \subseteq \text{Plays}_s(\mathcal{G})$ and $\text{Plays}_I^{\sigma}(\mathcal{G}) \subseteq \text{Plays}_I(\mathcal{G})$.

3 Responsibility in Non-cooperative Games

We now identify three qualitative notions of responsibility which we present in decreasing order of their logical strength.

3.1 Forward Responsibility

The individual responsibility of player i will be an average of the marginal contribution of i to *coalitional responsibility*, i.e., the responsibility of a group of players C. We first formalize that the players in C act *collaboratively*.

Definition 4 (Game induced by a coalition)**.** *Let $\mathcal{G}$ be an n-player game and $C \subseteq \{1, ..., n\}$. The 2-player game $\mathcal{G}_C$ is obtained from $\mathcal{G}$ as follows: The two players C and $\overline{C}$ are in control of the states in $P_C = \bigcup_{i \in C} P_i$ and $P_{\overline{C}} = \bigcup_{i \in \{1,...,n\} \setminus C} P_i$. Two states $s, s' \in P_C$ belong to the same information set in $\mathcal{G}_C$ if and only if they belong to the same information set in $\mathcal{G}$ and $\text{hist}_{\mathcal{G}}(s, C) = \text{hist}_{\mathcal{G}}(s', C)$. Similarly for $\overline{C}$. The labeling of terminal states remains unchanged, as do the states and distributions in control of player* `Nature`.

The information sets are the coarsest refinement of the existing information sets such that $\mathcal{G}_C$ is a game of perfect recall. The rationale for this is that the coalition C will share knowledge among its members; states that are indistinguishable by one player in C can become distinguishable by C because another player in the coalition can tell them apart.

Example 1 (Matching pennies). Two players independently choose heads or tails, and E is the event that they made opposite choices. This game is depicted in Figure 1, where the notation $s : i$ expresses that state s is in control of player i, dashed lines connect states in the same information set, and the red leaves indicate E-plays. If we did not refine the information sets in $\mathcal{G}_C$, $C = \{1, 2\}$ could not distinguish s_1 and s_2 even though a coalition member chose the action that produces the branches to these states. Thus enforcing perfect recall in $\mathcal{G}_C$ is necessary to model epistemic knowledge in the coalitional setting.

Definition 5 (Forward responsibility)**.** *Let $\mathcal{G}$ be an n-player game. We say that $C \subseteq \{1, ..., n\}$ is* forward responsible *(henceforth* **f***-responsible) for E if* **(F)** *there is a strategy σ of the player C in $\mathcal{G}_C$ such that all plays in $\text{Plays}^{\sigma}(\mathcal{G}_C)$ have label $\neg E$, and C is minimal with respect to property* **(F)**.

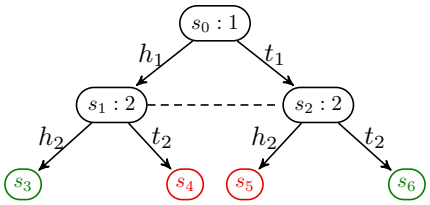

Figure 1: Matching pennies

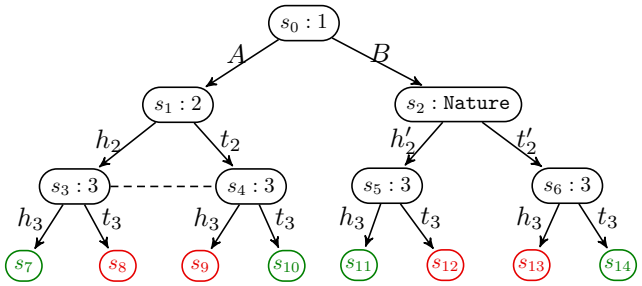

Figure 2: 3-player matching pennies

Being **f**-responsible means that C wins the game $\mathcal{G}_C$ under the reachability objective $\neg E$, and this condition is not satisfied for any proper subset of C. In order to illustrate Definition 5 and the subsequent notions of responsibility, we use the following toy scenario as running example. More interesting cases are postponed to the collection gathered in Section 4.

Example 2. Player 1 first decides between mode A and mode B. In mode A, players 2 and 3 independently choose a side of a coin as in Figure 1. In mode B, `Nature` tosses a coin, the result is revealed, and then player 3 chooses a side. The variable E denotes the event where the two sides of coins are not the same. See Figure 2 for a visualization of this game.

No single player is **f**-responsible. This is obvious for players 1 and 2, and for player 3 this is due to the non-trivial information set $\{s_3, s_4\}$: the strategy that enforces $\neg E$ in one of the states prevents it in the other. The coalition $\{1, 3\}$ is **f**-responsible since player 1 can choose to move into mode B, and player 3 responds according to what is revealed to him by `Nature`. Likewise $\{2, 3\}$ is **f**-responsible.

3.2 Strategic Backward Responsibility

In contrast to the preceeding section, backward responsibility is defined relative to a given play ρ ending in E. We distinguish between *strategic* backward responsibility and *causal* backward responsibility. Informally speaking, strategic backward responsibility means that C had the power to prevent E as the play ρ evolved *given its epistemic knowledge*:

Definition 6 (Strategic backward responsibility)**.** *The set C is* strategically backward responsible *(henceforth* **s***-responsible) for E based on an E-play ρ if* **(S)** *there exist a state s on ρ and a strategy σ of the player C in $\mathcal{G}_C$ such that:*

1. *ρ is consistent with σ until s is reached, and*
2. *all plays in $\text{Plays}_I^{\sigma}(\mathcal{G}_C)$ have label $\neg E$, where I is the information set of s,*

and C is minimal with respect to property **(S)**.

Clearly, each $\mathbf{f}$-responsible coalition, or even any coalition C satisfying property ($\mathbf{F}$) contains an $\mathbf{s}$-responsible coalition based on any E-play. In order to show property ($\mathbf{S}$) for C one can simply take s to be the initial state and take the strategy for C that globally enforces $\neg E$.

Example 3. Consider again our running example depicted in Figure 2. First let ρ be one of the two E-plays ending in s_8 and s_9. No single player is $\mathbf{s}$-responsible for E based on one of these plays. However, the coalitions $\{1,3\}$ and $\{2,3\}$ are $\mathbf{s}$-responsible for E based on ρ since they are $\mathbf{f}$-responsible. Based on the two E-plays ending in s_{12} and s_{13}, the single player 3 is strategically backwards responsible for E since he could have chosen differently in s_5, respectively, s_6.

Theorem 1 (Relating forward and backward responsibility)**.** *In a game of perfect recall, a coalition $C \subseteq \{1,...,n\}$ is $\mathbf{f}$-responsible for E if and only if it contains an $\mathbf{s}$-responsible coalition for E based on all E-plays, and is minimal with respect to this property.*

Remark 1. *Theorem 1 would not hold without the refinement of information sets in $\mathcal{G}_C$ illustrated in Example 1.*

3.3 Causal Backward Responsibility

A set of players might have caused E unknowingly and even inadvertently. For example, in the simple 2-player matching pennies scenario depicted in Figure 1, imagine that player 1 picks heads and player 2 picks tails. Both players may be held responsible for the result since changing their individual actions would have prevented it. However, *they were simply not aware of this fact*. The lack of knowledge that renders a strategic sense of responsibility implausible in this scenario refers to the uncertainty of the opponent's strategy (which also entails the uncertainty that comes with non-singleton information sets). Causal backward responsibility leverages this lack of knowledge by fixing a strategy profile for the opposing coalition. Intuitively, the coalition C *guesses* the strategy of $\overline{C}$ and thus attains the *hypothetical* knowledge to anticipate the actions of $\overline{C}$.

Definition 7 (Causal backward responsibility)**.** *Let σ be a strategy profile in $\mathcal{G}$ such that $\mathrm{Plays}^{\sigma}(\mathcal{G})$ contains an E-path ρ. Let $(\sigma_C, \sigma_{\overline{C}})$ be the strategy profile induced by σ in $\mathcal{G}_C$. Then $C \subseteq \{1,...,n\}$ is* causally backward responsible *(henceforth $\mathbf{c}$-responsible) for E based on ρ and σ if* ($\mathbf{C}$) *there exists a strategy σ'_C for C in $\mathcal{G}_C$ such that all plays in $\mathrm{Plays}^{\sigma'_C, \sigma_{\overline{C}}}(\mathcal{G}_C)$ that are also consistent with* `Nature`*'s random choices on ρ have label $\neg E$, and C is minimal with respect to property* ($\mathbf{C}$).

Intuitively, C satisfies condition ($\mathbf{C}$) if C's actions made a difference to the outcome when everything else is held fixed: The E-path ρ consistent with the strategy profile contains a state from which the coalition C *could* have employed a different strategy that enforces reaching $\neg E$ *provided that* C's guess on the opponent's strategy was right. Fixing strategies might seem like a far-fetched setup. We illustrate in Section 4 that many scenarios naturally determine a strategy profile, and we show in Section 5 that this canonically corresponds to fixing a *context* for a causal model.

Example 4. Consider once more the scenario depicted in Figure 2. In the following denote the Dirac distribution over $\mathrm{Act}(s)$ concentrated on a by $s \mapsto a$. Based on the strategy profile σ^1 given by $s_0 \mapsto B, s_1 \mapsto h_2, \{s_3, s_4\} \mapsto h_3, s_5 \mapsto t_3, s_6 \mapsto t_3$ and the play ending in s_{12}, players 1 and 3 are $\mathbf{c}$-responsible for E; player 1 can prevent E by switching to A since players 2 and 3 make identical choices in mode A, and player 3 can prevent E by choosing heads in s_5. The reader is invited to check that every single player is $\mathbf{c}$-responsible for E based on the strategy profile σ^2 given by $s_0 \mapsto A, s_1 \mapsto h_2, \{s_3, s_4\} \mapsto t_3, s_5 \mapsto h_3, s_6 \mapsto t_3$.

The example illustrates that $\mathbf{c}$-responsibility is an approach to identify 'causes' for E irrespective of epistemic information available to the players. A comparison to Halpern and Pearl's actual causes is given in Section 5.

Proposition 1 (From strategic responsibility to causal responsibility)**.** *Let σ be a strategy profile and let $\rho \in \mathrm{Plays}^{\sigma}(\mathcal{G})$ be an E-play. If C is $\mathbf{s}$-responsible for E based on ρ, then C contains a $\mathbf{c}$-responsible coalition for E based on ρ and σ.*

3.4 Quantifying Responsibility

All responsibility notions considered so far are qualitative. We now take the analysis one step further and *quantify* how much responsibility each *individual* player has. For this we employ the *Shapley value* for cooperative games [Shapley, 1953], i.e., games defined by a function $g\colon 2^n \to \mathbb{R}$, where $g(C)$ represents the common gain (or cost, depending on the situation) which coalition C can achieve collaboratively.

The responsibility notions $\{\mathbf{f}, \mathbf{s}, \mathbf{c}\}$ (i.e., $\mathbf{f}$orward, $\mathbf{s}$trategically backward, $\mathbf{c}$ausally backward) in an extensive form game naturally induce cooperative games. Let $\mathbf{t} \in \{\mathbf{f}, \mathbf{s}, \mathbf{c}\}$; if $\mathbf{t} = \mathbf{s}$ we assume that an E-play ρ is fixed, and if $\mathbf{t} = \mathbf{c}$, we further assume that a strategy profile σ is fixed (which we suppress from the notation for readability). We define the *induced* cooperative game $g_{\mathcal{G},\mathbf{t}}\colon 2^n \to \mathbb{R}$ by setting $g_{\mathcal{G},\mathbf{t}}(C)$ to be 1 if C contains a $\mathbf{t}$-responsible coalition for E (based on ρ, resp. σ) and to be 0 otherwise.

Definition 8 (Responsibility value)**.** *The* responsibility value $\mathrm{resp}_{\mathcal{G},\mathbf{t}}(i)$ *of player i in an n-player extensive form game $\mathcal{G}$ is the Shapley value of i in the induced cooperative game $g_{\mathcal{G},\mathbf{t}}$, i.e.,*

$$\mathrm{resp}_{\mathcal{G},\mathbf{t}}(i) = \frac{1}{n!} \sum_{\pi \in S_n} g_{\mathcal{G},\mathbf{t}}(\pi_{\geq i}) - g_{\mathcal{G},\mathbf{t}}(\pi_{\geq i} \setminus \{i\}),$$

where S_n denotes the set of permutations on $\{1,...,n\}$ and $\pi_{\geq i} = \{j \in \{1,...n\} \mid \pi(j) \geq \pi(i)\}$.

We have $\mathrm{resp}_{\mathcal{G},\mathbf{t}}(i) > 0$ if and only if player i belongs to a $\mathbf{t}$-responsible coalition, i.e., if i leaves the coalition, then it is not $\mathbf{t}$-responsible anymore. Moreover, we have $\sum_i \mathrm{resp}_{\mathcal{G},\mathbf{t}}(i) = \mathrm{resp}_{\mathcal{G},\mathbf{t}}(\{1,...,n\}) - \mathrm{resp}_{\mathcal{G},\mathbf{t}}(\emptyset)$ by a simple telescope sum argument, and this value is 1 unless $C = \emptyset$ is $\mathbf{t}$-responsible or $C = \{1,\ldots,n\}$ is not $\mathbf{t}$-responsible. Apart from these latter exceptional cases, the responsibility value measures the relative responsibility of each player against the other players within a given game. It is not intended to be compared across different types of responsibility, across plays, or even models.

The exceptional cases above are the only instances in which our theory allows *responsibility voids* [Braham and van Hees, 2018], i.e., $\mathrm{resp}_{\mathcal{G},\mathbf{t}}(i) = 0$ for every player i. This is not unintended: If $C = \emptyset$ is $\mathbf{t}$-responsible, then all plays must have label $\neg E$, meaning that (as far as the event E is concerned) the actions of the agents are *irrelevant*. If $C = \{1, \ldots, n\}$ is not $\mathbf{t}$-responsible, then ensuring $\neg E$ is *impossible*, and if the game does not involve randomization, this means that no play has label $\neg E$. We present a scenario involving responsibility voids in Example 8.

Example 5. In our running example of Figure 2 we have $\mathrm{resp}_{\mathcal{G},\mathbf{f}}(1) = \mathrm{resp}_{\mathcal{G},\mathbf{f}}(2) = 1/6$ and $\mathrm{resp}_{\mathcal{G},\mathbf{f}}(3) = 2/3$. Based on a play ending in s_8 or s_9, the same value arise for $\mathrm{resp}_{\mathcal{G},\mathbf{s}}(i)$. If ρ ends in s_{12} or s_{13}, then $\mathrm{resp}_{\mathcal{G},\mathbf{s}}(3) = 1$ and $\mathrm{resp}_{\mathcal{G},\mathbf{s}}(i) = 0$ for $i = 1, 2$. Based on the strategy profile σ^2 of Example 4 we have $\mathrm{resp}_{\mathcal{G},\mathbf{c}}(i) = 1/3$ for every player.

3.5 Computational Complexity

Remarkably, the complexity of deciding responsibility for a coalition is polynomial time. Our algorithm is based on the polynomial time procedure for solving two-player zero-sum games [Koller and Megiddo, 1992; von Stengel, 1996].

Theorem 2 (Complexity of deciding responsibility)**.** *Given an n-player game, a coalition $C \subseteq \{1, ..., n\}$, and $\mathbf{t} \in \{\mathbf{f}, \mathbf{s}, \mathbf{c}\}$. If $\mathbf{t} = \mathbf{s}$ we assume that an E-path ρ is fixed, and if $\mathbf{t} = \mathbf{c}$, we further assume that a strategy profile σ is fixed. It is decidable in polynomial time whether C is $\mathbf{t}$-responsible for E (based on ρ, resp., σ).*

A straightforward consequence of Theorem 2 is that deciding whether $\mathrm{resp}_{\mathcal{G},\mathbf{t}}(i) > 0$ belongs to NP, and computing $\mathrm{resp}_{\mathcal{G},\mathbf{t}}(i)$ belongs to #P.

4 Examples

We now illustrate our notion of responsibility allocation on several examples in the moral philosophy literature.

Example 6 (Bystanders). Consider a car accident with three victims in immediate need of first aid. There are four bystanders $1, 2, 3, 4$ who arrive in this order and who can help one of the victims (we implicitly assume that if one victim is already being helped, and another bystander decides to help, then she helps an unassisted victim). Let E be the event where at least one victim dies. Then a coalition is $\mathbf{f}$-responsible if it contains exactly three players, and we get $\mathrm{resp}_{\mathcal{G},\mathbf{f}}(i) = 1/4$, which just means that they have globally the same power. Consider the play where bystanders 1 and 3 help and the others do not help. Then the only $\mathbf{s}$-responsible coalition is $\{4\}$. Hence $\mathrm{resp}_{\mathcal{G},\mathbf{s}}(4) = 1$, which is a numerical interpretation of the fact that 4 *knows* that she can (or must) make the difference for the remaining victim to survive, while player 2 does not have this certainty.

Next consider the scenario in which 1 helps, but the others do not help irrespective of any previous actions. Then $\mathrm{resp}_{\mathcal{G},\mathbf{c}}(1) = 0$, $\mathrm{resp}_{\mathcal{G},\mathbf{c}}(i) = 1/3$ for $i = 2, 3, 4$, which explains that the remaining bystanders are equally responsible for the death of the victims. Finally imagine that bystander 4 would have helped if bystander 3 had helped (the *bystander effect*), but 2 and 3 never help. In this scenario we get $\mathrm{resp}_{\mathcal{G},\mathbf{c}}(1) = 0, \mathrm{resp}_{\mathcal{G},\mathbf{c}}(2) = \mathrm{resp}_{\mathcal{G},\mathbf{c}}(4) = 1/6$, and $\mathrm{resp}_{\mathcal{G},\mathbf{c}}(3) = 2/3$ which puts more causal responsibility on 3's shoulders for not being an exemplar to 4.

Example 7 (Marksmen). Ten marksmen form a firing squad for the execution of a prisoner. They know that exactly one of them has a live bullet in his rifle, but they do not know which one. All ten have (at least theoretically) the choice of firing or not. This is modeled as a game where `Nature` first chooses the player i^* that has the live bullet and then the marksmen concurrently shoot or not. Let E be the event that the prisoner dies, and consider an E-play ρ (i.e., i^* shoots). Due the uncertainty introduced by `Nature`, only the coalition consisting of all marksmen is $\mathbf{f}$-responsible, respectively, $\mathbf{s}$-responsible for E based on ρ, and hence $\mathrm{resp}_{\mathcal{G},\mathbf{f}}(i) = \mathrm{resp}_{\mathcal{G},\mathbf{s}}(i) = 1/10$. However, we always have $\mathrm{resp}_{\mathcal{G},\mathbf{c}}(i^*) = 1$ independent of the strategies of the other nine marksmen – i.e. the one with the live bullet carries the entire causal responsibility.

Example 8 (Bogus prevention). A person P is protected by a bodyguard B, who suspects a poisonous attack and puts an antidote into P's coffee. Indeed, an assassin A tries to poison P, but the poison is neutralized by the antidote. Consider the event E that P survives. Based on the play ρ and strategies described above, no single player is $\mathbf{f}$- or $\mathbf{s}$-responsible, but B is $\mathbf{c}$-responsible, while A is not. Now consider the alternative model in which A, seeing B's action, poisons the coffee exactly if B chose to put in the antidote, which determines a strategy σ_A (e.g., because A and B are old friends, and A wants B to be considered important by P). Suppose that B puts in the antidote, thus determining a strategy σ_B (inducing the same path ρ as above). Based on these strategies no individual player is $\mathbf{c}$-responsible for P's survivial! We emphasize this point since all versions of Halpern-Pearl's theory *do* consider B putting in the antidote an actual cause of P's survival, even though the structural equations determined by this model and context already preclude P's death.

5 Comparison to Other Approaches

We now provide a detailed comparison of our notions with related approaches to responsibility allocation.

Causal Models

An influential formal concept of causality is Halpern and Pearl's notion of *actual causes* in *causal* models [Halpern and Pearl, 2005]. In a nutshell, a causal model $\mathcal{M}$ consists of variables $\mathcal{U} \cup \mathcal{V}$ and functions F_X that determine the values for the variables $X \in \mathcal{V}$ depending on the other variables. A *context* for $\mathcal{M}$ is a tuple $\vec{u}$ specifying values for $\mathcal{U}$, which induce unique values for $\mathcal{V}$. A *causal formula* is of the form $[\vec{Y} \leftarrow \vec{y}]\varphi$, where φ is a Boolean combination of statements about the values of the variables in $\mathcal{V}$, and $[\vec{Y} \leftarrow \vec{y}]$ denotes the act of forcing $\vec{Y} \subseteq \mathcal{V}$ to have values $\vec{y}$. This *intervention* may or may not affect the values of other variables in $\mathcal{V}$. If φ becomes a true statement upon evaluating all variables according to context $\vec{u}$, an intervention $[\vec{Y} \leftarrow \vec{y}]$, and the functions F_X, then we write $(\mathcal{M}, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}]\varphi$.

A *but-for cause* for φ is a minimal set of variables $\vec{X} \subseteq \mathcal{V}$ such that $(\mathcal{M}, \vec{u}) \models (\vec{X} = \vec{x})$, $(\mathcal{M}, \vec{u}) \models \varphi$, and there exist

alternative values $\vec{x}'$ for $\vec{X}$ such that $(\mathcal{M}, \vec{u}) \models [\vec{X} \leftarrow \vec{x}']\neg\varphi$. The intuition is that *but for* the variables in $\vec{X}$, the formula φ would not hold. In fact, [Halpern and Pearl, 2005] define a larger class of *actual causes* for which the axioms above are generalized to allow for *witnesses* of the but-for condition. In [Baier *et al.*, 2021] we present a translation of a causal model $\mathcal{M}$ to an extensive form game $\mathcal{G}(\mathcal{M})$ with players $\mathcal{V}$ such that a formula φ induces a labeling $E(\varphi)$ of $\mathcal{G}(\mathcal{M})$ and a context $\vec{u}$ naturally induces a path $\rho_{\vec{u}}$ and a strategy profile $\sigma^{\vec{u}}$. Then:

Theorem 3. *For a causal model $\mathcal{M}$ with context $\vec{u}$, $\vec{X} \subseteq \mathcal{V}$ is* **c***-responsible for $E(\varphi)$ based on $\rho_{\vec{u}}$ and $\sigma^{\vec{u}}$ in $\mathcal{G}(\mathcal{M})$ iff $\vec{X} = \vec{x}$ is a but-for cause for φ in $(\mathcal{M}, \vec{u})$.*

It is arguably a feature of Halpern and Pearl's framework that it can ascribe actual causality to a larger class of events than but-for causes, e.g., it can distinguish the two players in the Suzy-Billie rock-throwing example from the introduction. It is noteworthy, though, that in this case (and many others, cf. [Halpern, 2015]) the structural model approach distinguishes Suzy as the sole actual cause based on *auxiliary variables* that are *not* actions of independent players. In the rock-throwing model the witness is Billy's rock not hitting the bottle rather than Billy's throwing. But the fact that Billy's rock did not hit the bottle is not a decision made by Billy—it just happened on the basis of previous actions as dictated by the structural equations. Auxiliary variables do not entail *agency* nor the possibility to *act* otherwise (compare the introduction). Our intention is to locate responsibility exclusively within and against (the actions of) autonomous players.

Degree of Responsibility and Blame

Chockler and Halpern assign to actual causes (in the original definition [Halpern and Pearl, 2005]) a *degree of responsibility* which tries to measure how crucial the cause is to the effect [Chockler and Halpern, 2004]. It is essentially defined as the inverse of the size of a minimal set of interventions such that swapping the cause's value flips the truth value of the effect. It is noteworthy that the degree of responsibility is designed to be comparable across models, while the responsibility value measures the responsibility of each player against the other players within a given model.

Translated to our approach, the degree of $\mathbf{t}$-responsibility of a player would be the inverse of the size of the smallest $\mathbf{t}$-responsible coalition that she belongs to. In our running example, all three players would have degree of $\mathbf{f}$-responsibility $1/2$, since $\{1, 3\}$ and $\{2, 3\}$ are $\mathbf{f}$-responsible. However, the fact that 3 belongs to both is ignored. Our responsibility value, on the other hand, takes *all* $\mathbf{t}$-responsible coalitions into account in which the player participates. This captures the idea that belonging to many responsible coalitions makes the player less dependent on others and hence more powerful.

Concurrent Game Structures

The work closest to ours is [Yazdanpanah *et al.*, 2019], where notions of forward and backward responsibility are defined in the context of concurrent game structures. However, there are several differences: First, they do not consider a causal notion of backward responsibility, and their backward responsibility does not take the epistemic state appropriately into account: An alternative strategy of a responsible coalition is required to bring about an alternative outcome only from a single state on the play, not from *all* states in the same information set. For example, in their Figure 2, the second player E_2 is backward responsible based on the play $q_0q_1q_4$ (he can avoid the red states from q_1), but not based on the play $q_0q_2q_6$, even though he acted in *exactly* the same way based on *exactly* the same knowledge. Second, their notion of backward responsibility is asymmetric: In a game with two players choosing simultaneously and independently from the same set of actions based on the same knowledge, it can happen that they perform identical actions, but one is backward responsible and the other one is not. Third, while [Yazdanpanah *et al.*, 2019, Theorem 3.3] states that backward responsibility is equivalent to forward responsibility from *all states* on a play, our Theorem 1 states that forward responsibility is equivalent to strategic backward responsibility on *all plays*. This is a crucial shift of perspective that has—to the best of our knowledge—not been proved before for similar notions of responsibility. In fact, it responds to a central philosophical question left open in prior work: the relation between *general* (or type-level) responsibility and *specific* (or token-level) responsibility.

Proof-theoretic Approaches

The work [Broersen, 2011] as well as the series of papers [Naumov and Tao, 2020a; Naumov and Tao, 2020b; Naumov and Tao, 2020c] provide proof-theoretic approaches to responsibility (there called *blameworthiness*, a term we avoid as it typically involves normative features). They define modal logics with various forms of responsibility modalities and sound and complete axiomatizations for the logical systems. Instead, we take an operational, model-theoretic approach. The computational complexity of their formalizations is left open. Given the semantics of the blameworthiness modalities of [Naumov and Tao, 2020a; Naumov and Tao, 2020b; Naumov and Tao, 2020c], one can expect at least PSPACE-hardness (or even undecidability) for checking blameworthiness in these formalisms. Known complexity results about (original or Chellas's) STIT formulas also point to PSPACE-hardness for the notions of [Broersen, 2011]. The above papers do not provide a *quantitative* version of responsibility assigned to *individuals* or a way to compare responsibilities.

6 Conclusion

We have argued that extensive form games provide a conceptually convenient formal framework for responsibility ascription with just the right trade-off between expressiveness and tractability. We have defined qualitative (coalitional) and quantitative (individual) responsibility notions. Through a set of examples, we demonstrate that our notions capture intuitive responsibility ascription in many subtle examples.

Acknowledgements

This work was funded by DFG grant 389792660 as part of TRR 248 – CPEC, the Cluster of Excellence EXC 2050/1 (CeTI, project ID 390696704, as part of Germany's Excellence Strategy), DFG-projects BA-1679/11-1 and BA-1679/12-1, and the European Research Council under the Grant Agreement 610150 (ERC Synergy Grant ImPACT)

References

[Aleksandrowicz *et al.*, 2014] Gadi Aleksandrowicz, Hana Chockler, Joseph Y. Halpern, and Alexander Ivrii. The Computational Complexity of Structure-Based Causality. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 974–980. AAAI Press, 2014.

[Baier *et al.*, 2021] Christel Baier, Florian Funke, and Rupak Majumdar. A Game-Theoretic Account of Responsibility Allocation. Technical Report arXiv:2105.09129, 2021.

[Beer *et al.*, 2009] Ilan Beer, Shoham Ben-David, Hana Chockler, Avigail Orni, and Richard Trefler. Explaining counterexamples using causality. In Ahmed Bouajjani and Oded Maler, editors, *Computer Aided Verification*, pages 94–108. Springer Berlin Heidelberg, 2009.

[Braham and van Hees, 2012] Matthew Braham and Martin van Hees. An Anatomy of Moral Responsibility. *Mind*, 121 (483):601–634, 2012.

[Braham and van Hees, 2018] Matthew Braham and Martin van Hees. Voids or Fragmentation: Moral Responsibility For Collective Outcomes. *The Economic Journal*, 128 (602):F95–F113, 2018.

[Broersen, 2011] Jan Broersen. Deontic epistemic stit logic distinguishing modes of mens rea. *Journal of Applied Logic*, 9(2):137–152, 2011.

[Bulling and Dastani, 2013] Nils Bulling and Mehdi Dastani. Coalitional Responsibility in Strategic Settings. In *Computational Logic in Multi-Agent Systems*, pages 172–189, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[Chockler and Halpern, 2004] Hana Chockler and Joseph Y. Halpern. Responsibility and Blame: A Structural-Model Approach. *J. Artif. Int. Res.*, 22(1):93–115, October 2004.

[Chockler *et al.*, 2008] Hana Chockler, Joseph Y. Halpern, and Orna Kupferman. What causes a system to satisfy a specification? *ACM Transactions on Computational Logic*, 9(3):20:1–20:26, 2008.

[Chockler, 2016] Hana Chockler. Causality and Responsibility for Formal Verification and Beyond. In *Proceedings of the First Workshop on Causal Reasoning for Embedded and safety-critical Systems Technologies, CREST*, volume 224 of *EPTCS*, pages 1–8, 2016.

[Friedenberg and Halpern, 2019] Meir Friedenberg and Joseph Y. Halpern. Blameworthiness in Multi-Agent Settings. In *The 33rd AAAI Conf. on Artificial Intelligence, AAAI 2019*, pages 525–532. AAAI Press, 2019.

[Halpern and Pearl, 2005] Joseph Y. Halpern and Judea Pearl. Causes and Explanations: A Structural-Model Approach. Part I: Causes. *The British Journal for the Philosophy of Science*, 56(4):843–887, 2005.

[Halpern, 2015] Joseph Y. Halpern. A Modification of the Halpern-Pearl Definition of Causality. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 3022–3033. AAAI Press, 2015.

[Halpern, 2016] Joseph Y. Halpern. *Actual Causality*. MIT Press, 2016.

[Hart and Honoré, 1959] H. L. A. Hart and A. M. Honoré. *Causation in the law*. Oxford University Press, 1959.

[Koller and Megiddo, 1992] Daphne Koller and Nimrod Megiddo. The complexity of two-person zero-sum games in extensive form. *Games and Economic Behavior*, 4(4):528 – 552, 1992.

[Kuhn, 1953] Harold W. Kuhn. *Extensive Games and the Problem of Information*, pages 193 – 216. Princeton University Press, Princeton, 1953.

[Lundberg and Lee, 2017] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS 2017*, 2017.

[Naumov and Tao, 2020a] Pavel Naumov and Jia Tao. Blameworthiness in security games. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(03):2934–2941, Apr. 2020.

[Naumov and Tao, 2020b] Pavel Naumov and Jia Tao. Duty to Warn in Strategic Games. In *Proceedings of the 19th International Conference on Autonomous Agents and Multi-Agent Systems*, AAMAS '20, page 904–912, 2020.

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

[Owen, 1995] Guillermo Owen. *Game Theory*. Academic Press, 1995.

[Scanlon, 1998] Thomas M. Scanlon. *What we owe to each other*. Cambridge University Press, 1998.

[Shapley, 1953] Lloyd S. Shapley. A value for n-person games. In *Contributions to the Theory of Games. Vol. II.*, pages 307–317. Princeton University Press, 1953.

[van de Poel, 2011] Ibo van de Poel. The Relation Between Forward-Looking and Backward-Looking Responsibility. In *Moral Responsibility: Beyond Free Will and Determinism*, pages 37–52. Springer Netherlands, Dordrecht, 2011.

[von Neumann, 1928] John von Neumann. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, 100:295–320, 1928.

[von Stengel, 1996] Bernhard von Stengel. Efficient computation of behavior strategies. *Games and Economic Behavior*, 14(2):220 – 246, 1996.

[Štrumbelj and Kononenko, 2014] Erik Štrumbelj and Igor Kononenko. Explaining Prediction Models and Individual Predictions with Feature Contributions. *Knowl. Inf. Syst.*, 41(3):647–665, December 2014.

[Yazdanpanah and Dastani, 2016] Vahid Yazdanpanah and Mehdi Dastani. Distant group responsibility in multi-agent systems. In *PRIMA 2016: Principles and Practice of Multi-Agent Systems*, pages 261–278, Cham, 2016. Springer International Publishing.

[Yazdanpanah *et al.*, 2019] Vahid Yazdanpanah, Mehdi Dastani, Wojciech Jamroga, Natasha Alechina, and Brian Logan. Strategic Responsibility Under Imperfect Information. In *The 18th Intern. Conf. on Autonomous Agents and MultiAgent Systems*, AAMAS '19, page 592–600, 2019.