On Blame Attribution for Accountable Multi-Agent Sequential Decision Making

Stelios Triantafyllou
MPI-SWS
strianta@mpi-sws.org

Adish Singla
MPI-SWS
adishs@mpi-sws.org

Goran Radanovic
MPI-SWS
gradanovic@mpi-sws.org

Abstract

Blame attribution is one of the key aspects of accountable decision making, as it provides means to quantify the responsibility of an agent for a decision making outcome. In this paper, we study blame attribution in the context of cooperative multi-agent sequential decision making. As a particular setting of interest, we focus on cooperative decision making formalized by Multi-Agent Markov Decision Processes (MMDPs), and we analyze different blame attribution methods derived from or inspired by existing concepts in cooperative game theory. We formalize desirable properties of blame attribution in the setting of interest, and we analyze the relationship between these properties and the studied blame attribution methods. Interestingly, we show that some of the well known blame attribution methods, such as Shapley value, are not performance-incentivizing, while others, such as Banzhaf index, may over-blame agents. To mitigate these value misalignment and fairness issues, we introduce a novel blame attribution method, unique in the set of properties it satisfies, which trade-offs explanatory power (by under-blaming agents) for the aforementioned properties. We further show how to account for uncertainty about agents' decision making policies, and we experimentally: a) validate the qualitative properties of the studied blame attribution methods, and b) analyze their robustness to uncertainty.

... a body of people[1]*, holding themselves accountable to nobody, ought not to be trusted by anybody.*

—Thomas Paine, A philosopher and a political activist.

1 Introduction

With the widespread usage of artificial intelligence (AI) in everyday life [1, 2, 3], accountability has become one of the central problems in the study of AI. Much recent research studied what constitutes accountability in the context of AI and how to design accountable AI systems [4, 5, 6], and recent policies and legislations [7] are increasingly highlighting the importance of accountability, aiming to provide guidelines for developing and deploying accountable AI systems.

Accountability is a relatively broad term, and it typically involves an actor (or multiple actors) justifying their decisions and facing consequences for actions taken [6, 8]. Hence, two critical aspects of accountability are explainability and blame attribution. Recent work proposed various methods for explaining, interpreting, understanding, and certifying algorithmic decision-making and its outcomes [9, 10, 11, 12, 13, 14]. In this paper we study the other critical aspect of accountability – *blame attribution*.

[1] Originally, and by modern standards outdated, Thomas Paine used phrasing with the word *men*.

35th Conference on Neural Information Processing Systems (NeurIPS 2021)

In multi-agent decision making, one of the central roles of blame attribution is assigning blame for undesirable outcomes or, broadly speaking, for the system's inefficiency. Prior work on responsibility and blame in AI [15, 16, 17, 18] has recognized some of the core challenges in attributing blame, including the fact that disentangling agents' contributions to the final outcome is not a trivial task. Such challenges are particularly prominent in sequential settings where past decisions influence the future ones [16].

In this paper, we consider the task of allocating a score to an agent, which represents the degree of its blame, and reflects its contributions to the total inefficiency of the multi-agent system. We focus on cooperative sequential decision making, formalized by multi-agent Markov decision processes (MMDPs) [19], where the outcome of interest is the expected discounted return of the agents' joint policy. Concretely, given an MMDP and the agents' joint policy (true or estimated), we ask: *How to score each agent so that the agents' scores satisfy desirable properties?*

To answer this question, we turn to cooperative game theory and consider blame attribution methods that are derived from or inspired by existing concepts in the cost sharing, data valuation, and coalition formation literature [20, 21, 22, 23, 24, 25, 26, 27], such as core [28], Shapley value [29, 30], or Banzhaf index [31, 32]. Taking this perspective on blame attribution, we study blame attribution for accountable multi-agent sequential decision making. More concretely:

- We formalize desirable properties that blame attribution methods should satisfy in cooperative multi-agent sequential decision making. We identify properties that are typically not considered in the cost-sharing literature, yet are important for decision making. In particular, we introduce two novel properties: a) *performance monotonicity*, which states that, having fixed all the other agents to their policies, the blame assigned to an agent should not increase if the agent adopts a policy that results in a higher expected discounted return (implying that the method is performance-incentivizing); b) *Blackstone consistency*,[2] which states that an agent should not receive a higher blame just because the agents' policies are not fully known to the blame attribution procedure.
- We characterize the properties of the studied blame attribution methods. We show that some blame assignment methods, such as, Shapley value, are not performance-monotonic (and, hence, performance-incentivizing), while others, such as Banzhaf index, may over-blame agents. Motivated by these results, we introduce a novel blame attribution method that trade-offs explanatory power (by under-blaming agents) for the aforementioned properties.
- We provide algorithms for making the studied blame attribution methods Blackstone-consistent when the agents' policies are estimated. We also characterize the effect of uncertainty on blame attribution methods.
- Using a simulation-based testbed, we experimentally analyze the studied blame attribution methods, their qualitative properties, as well as their robustness to uncertainty. The experiments showcase the importance of the robustness considerations we study and indicate that typically more efficient blame attribution methods (i.e., those that assign more blame in total) are less robust to uncertainty.

1.1 Other Related Work

Apart from the works mentioned in the previous paragraphs, our work relates to different areas of moral philosophy, law, and AI, and here we highlight some of the most relevant references. Research in moral philosophy and law has extensively studied the problem of blame attribution, both in terms of human actors [34, 35, 36], as well as AI actors [37, 38, 39, 40]. We take some of the well known principles in moral philosophy and law in determining properties relevant for blame attribution, e.g., Blackstone consistency is inspired by Blackstone's ratio [33]. In AI, blame attribution has been studied through a more formal lens, utilizing causality [15, 16, 17] and/or game theory [18, 41], and primarily focusing on nuances related to defining notions and degrees of responsibility, blame, and blameworthiness. In contrast, we focus on cooperative sequential decision making, and analyze how different blame attribution methods from cooperative game theory fare under different blame attribution properties. Finally, our work is generally related to the *credit assignment* problem [42, 43], and more specifically to the credit assignment problem in multi-agent reinforcement learning [44, 45, 46]. However, our focus is not on supporting the learning processes of agents by reducing computational and statistical challenges of learning, but on evaluating the agents' contributions to the system's inefficiency, ideally in a fair and interpretable manner.

[2]This property is inspired by Blackstone's ratio: "It is better that ten guilty persons escape than that one innocent suffer" [33].

2 Formal Setting

In this section, we describe our formal setting, based on multi-agent Markov decision processes (MMDPs), and we formally model the blame attribution problem in sequential decision making. This section also introduces a set of desirable formal properties of blame attribution methods.

2.1 Preliminaries

We consider a cooperative multi-agent setting, formalized as a class of MMDPs $\mathcal{M}$ with n agents $\{1, ..., n\}$. Each MMDP in this class is a tuple $M = (\mathcal{S}, \{1, ..., n\}, \mathcal{A}, R, P, \gamma, \sigma)$ [19], where: $\mathcal{S}$ is the state space; $\mathcal{A} = \times_{i=1}^{n} \mathcal{A}_i$ is the action space, with $\mathcal{A}_i$ being the action space of agent i; R is the reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ specifying the reward obtained when agents $\{1, ..., n\}$ take a joint action; P specifies transitions with $P(s, a, s')$ denoting the probability of transitioning to s' from s when agents $\{1, ..., n\}$ take joint action $a = (a_1, ..., a_n)$; γ is the discount factor; and σ is the initial state distribution. $\mathcal{S}$ and $\mathcal{A}$ are finite and discrete. A (stationary) joint policy π is a mapping $\pi : \mathcal{S} \to \mathcal{D}(\mathcal{A})$, where $\mathcal{D}(\mathcal{A})$ is a probability simplex over $\mathcal{A}$, with $\pi(a|s)$ denoting the probability of taking joint action a in s. We assume that a joint policy π is factorizable into agents' policies, π_i, i.e., $\pi(a|s) = \pi_1(a_1|s) \cdots \pi_n(a_n|s)$. Therefore, we can define an agent i's policy π_i as a mapping from states to a distribution of agent i's actions, i.e., $\pi_i : \mathcal{S} \to \mathcal{D}(\mathcal{A}_i)$. We denote the set of all policies by $\Pi = \times_{i=1}^{n} \Pi_i$. We also define a standard performance measure. The expected discounted return of a joint policy π is defined as $J(\pi) = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} R(s_t, a_t) | s_1 \sim \sigma, \pi\right]$, where the initial state s_1 is sampled from σ, and the state-joint action pair of time-step t, (s_t, a_t), is obtained by executing joint policy π. We abuse our notation by denoting $J(\pi'_i, \pi_{-i}) = J(\pi_1, ..., \pi'_i, ..., \pi_n)$. Similarly, $J(\pi'_S, \pi_{-S}) = J(\pi'')$ for some $S \subseteq \{1, ..., n\}$, where $\pi''_i = \pi_i$ if $i \notin S$ and $\pi''_i = \pi'_i$ if $i \in S$.

2.2 Blame Attribution

Our goal is to assign blame to agents for failing to jointly achieve optimal performance. Given the agents' behavior policy, denoted by π^b, the inefficiency of the considered multi-agent system can be defined as $\Delta = J(\pi^*) - J(\pi^b)$, where $\pi^* \in \arg\max_\pi J(\pi)$ is an optimal joint policy. Similarly, we define the marginal inefficiency of a subset of agents $S \subseteq \{1, ..., n\}$ as $\Delta_S = J(\pi_S^{*|\pi^b}, \pi^b{}_{-S}) - J(\pi^b)$, as well as the marginal inefficiency of an agent i as $\Delta_i = J(\pi_i^{*|\pi^b}, \pi^b{}_{-i}) - J(\pi^b)$, where $\pi_S^{*|\pi^b}$ (resp. $\pi_i^{*|\pi^b}$) denotes an optimal policy of S (resp. i) assuming all other policies are fixed, i.e., $\pi_S^{*|\pi^b} \in \arg\max_{\pi_S} J(\pi_S, \pi^b{}_{-S})$ (resp. $\pi_i^{*|\pi^b} \in \arg\max_{\pi_i} J(\pi_i, \pi^b{}_{-i})$). A blame attribution method is a mapping $\Psi : \mathcal{M} \times \Pi \to \mathbb{R}^n_{\geq 0}$, where $\Psi(M, \pi^b)$ distributes blame for inefficiency Δ by assigning score β_i to agent i. The output of Ψ, i.e., the blame assignment, is denoted by β.

Uncertainty considerations. Since the agents' behavior policy π^b might not be known to the blame attribution procedure, we also define blame attribution under uncertainty as a mapping $\widehat{\Psi} : \mathcal{M} \times \mathcal{P}(\Pi) \to \mathbb{R}^n_{\geq 0}$ that outputs a blame assignment estimate $\widehat{\beta}$. Here, $\mathcal{P}(\Pi)$ represents a set whose elements express the knowledge about π^b. Inspired by the literature on robust MDPs [47, 48, 49], we encode such knowledge with uncertainty sets $\mathcal{P}(\pi^b)$, one associated to each state s, $\mathcal{P}(\pi^b, s)$, defined by the set of probability measures on $\mathcal{A}$. We assume that $\mathcal{P}(\pi^b)$ is consistent with π^b, i.e., $\pi^b(\cdot|s)$ is in $\mathcal{P}(\pi^b, s)$,[3,4] and that every $\pi(\cdot|s)$ in $\mathcal{P}(\pi^b, s)$ factorizes to $\pi(a|s) = \pi_1(a_1|s) \cdots \pi_n(a_n|s)$. Therefore, $\mathcal{P}(\pi^b, s)$ identifies the set of plausible stochastic actions that agent i takes in state s.

2.3 Desirable Properties

Our goal is to specify functions Ψ and $\widehat{\Psi}$, such that the blame assignments β and $\widehat{\beta}$ satisfy desirable properties. In the following text we denote these properties by $\mathcal{R}$. Below, we define properties that are taken from or inspired by the game theory literature [26, 50, 21], but translated to our setting[5]:

- *Validity* ($\mathcal{R}_V$): We say that Ψ is valid if it never distributes more blame than the observed inefficiency Δ. More formally, Ψ satisfies $\mathcal{R}_V$ (resp. ϵ-$\mathcal{R}_V$) if for every M and π^b, $\sum_{i=1}^{n} \beta_i \leq \Delta$ (resp. $\sum_{i=1}^{n} \beta_i \leq \Delta + \epsilon$), where $\beta = \Psi(M, \pi^b)$.

[3] Such $\mathcal{P}(\pi^b)$ can be derived from data containing agents' trajectories and be based on confidence intervals.
[4] $\pi^b(\cdot|s)$ could be in $\mathcal{P}(\pi^b, s)$ w.h.p., provided Blackstone consistency in Section 2.3 is similarly adjusted.
[5] Note that the terminology is slightly different.

- *Efficiency* ($\mathcal{R}_E$): A more strict condition is that the total distributed blame is equal to Δ. That is, Ψ satisfies $\mathcal{R}_E$ (resp. ϵ-$\mathcal{R}_E$) if for every M and π^b, $\sum_{i=1}^{n} \beta_i = \Delta$ (resp. $|\sum_{i=1}^{n} \beta_i - \Delta| \leq \epsilon$), where $\beta = \Psi(M, \pi^b)$.
- *Rationality* ($\mathcal{R}_R$): Similar to validity is rationality, which requires that blame distributed to any subset of agents S is not greater than Δ_S. That is, Ψ satisfies $\mathcal{R}_R$ (resp. ϵ-$\mathcal{R}_R$) if for every M, π^b, and $S \subseteq \{1, ..., n\}$, $\sum_{i \in S} \beta_i \leq \Delta_S$ (resp. $\sum_{i \in S} \beta_i \leq \Delta_S + \epsilon$), where $\beta = \Psi(M, \pi^b)$.
- *Symmetry* ($\mathcal{R}_S$): We say that Ψ is symmetric if it treats equal agents equally, i.e., agents that equally contribute to the inefficiency should receive the same blame. More formally, Ψ satisfies $\mathcal{R}_S$ (resp. ϵ-$\mathcal{R}_S$) if for every M and π^b, $\beta_i = \beta_j$ (resp. $|\beta_i - \beta_j| \leq \epsilon$) whenever $\Delta_{S \cup \{i\}} = \Delta_{S \cup \{j\}}$ for all $S \subseteq \{1, ..., n\} \backslash \{i, j\}$, where $\beta = \Psi(M, \pi^b)$.
- *Invariance* ($\mathcal{R}_I$): We say that Ψ is invariant if it assigns zero blame to agents who do not marginally contribute to inefficiency. More formally, Ψ satisfies $\mathcal{R}_I$ (resp. ϵ-$\mathcal{R}_I$) if for every M and π^b, $\beta_i = 0$ (resp. $\beta_i \leq \epsilon$) whenever $\Delta_{S \cup \{i\}} = \Delta_S$ for all S, where $\beta = \Psi(M, \pi^b)$.

Note that $\epsilon > 0$ in the definitions of ϵ-$\mathcal{R}$, and that we use these properties in our characterization result for blame attribution under uncertainty. Additionally, we consider two properties that relate the blame attribution output to the MMDP structure and the agents' behavior policies.

- *Contribution monotonicity* ($\mathcal{R}_{CM}$)[51]: We say that Ψ is contribution-monotonic if the blame it assigns to an agent depends only on its marginal contributions and monotonically so. More formally, Ψ satisfies $\mathcal{R}_{CM}$ (resp. ϵ-$\mathcal{R}_{CM}$) if for every two $(M^1, {\pi^b}^1)$ and $(M^2, {\pi^b}^2)$, ${\beta_i}^1 \geq {\beta_i}^2$ (resp. ${\beta_i}^1 \geq {\beta_i}^2 - \epsilon$) whenever $\Delta^1_{S \cup \{i\}} - \Delta^1_S \geq \Delta^2_{S \cup \{i\}} - \Delta^2_S$ for all S, where $\beta^1 = \Psi(M^1, {\pi^b}^1)$ and $\beta^2 = \Psi(M^2, {\pi^b}^2)$.
- *Performance monotonicity* ($\mathcal{R}_{PerM}$): We say that Ψ is performance-monotonic if it does not assign greater blame to agent i for adopting a policy that results in an equal or higher performance, assuming the other agents' policies fixed. More formally, consider any MMDP M, and any ${\pi^b}_{-i}$, π_i and π'_i such that $J(\pi_i, {\pi^b}_{-i}) \leq J(\pi'_i, {\pi^b}_{-i})$. We say that Ψ satisfies $\mathcal{R}_{PerM}$ (resp. ϵ-$\mathcal{R}_{PerM}$) if $\beta_i \geq \beta'_i$ (resp. $\beta_i \geq \beta'_i - \epsilon$) where $\beta = \Psi(M, (\pi_i, {\pi^b}_{-i}))$ and $\beta' = \Psi(M, (\pi'_i, {\pi^b}_{-i}))$.

The above definitions directly extend to $\widehat{\Psi}$ except that we require them to hold for all $\mathcal{P}(\pi_b)$. Additionally, we identify the following property for $\widehat{\Psi}$:

- *Blackstone consistency* ($\mathcal{R}_{BC}$): We say that $\widehat{\Psi}$ is Blackstone-consistent with Ψ if it never attributes more blame to an agent than Ψ. More formally, $\widehat{\Psi}$ satisfies $\mathcal{R}_{BC}(\Psi)$ if for any M, π^b and $\mathcal{P}(\pi^b)$, $\widehat{\beta}_i \leq \beta_i$, where $\beta = \Psi(M, \pi^b)$ and $\widehat{\beta} = \widehat{\Psi}(M, \mathcal{P}(\pi^b))$.

3 Game-Theoretic Approaches to Blame Attribution

In this section, we study blame attribution methods based on well known game theoretic notions, such as the core [28], Shapley value [29, 30], or Banzhaf index [31, 32]. We also introduce a novel blame attribution method, unique in the set of properties it satisfies. The proofs of our results can be found in Appendices G, H, and I.

3.1 Max-Efficient Rationality

We start with a relatively simple blame assignment method that puts rationality as a strict condition, and maximizes the efficiency of blame assignment under this constraint. We call this blame assignment method *max-efficient rationality*. More formally, max-efficient rationality can be defined via the following linear program:

$$\Psi_{MER}(M, \pi^b) := \max_{\beta} \sum_{i=1}^{n} \beta_i \qquad \text{s.t.} \qquad \sum_{i \in S} \beta_i \leq \Delta_S \quad \forall S \subseteq \{1, ..., n\}, \tag{P1}$$

where Δ_S are precomputed. Max-efficient rationality is inspired by the notion of core, but unlike the core, max-efficient rationality does not require $\mathcal{R}_E$ (efficiency) to hold. It is easy to show that the following properties are satisfied by any optimizer of (P1), Ψ_{MER}.

Proposition 1. *Every solution to the optimization problem* (P1)*, i.e., Ψ_{MER}, satisfies $\mathcal{R}_V$ (validity), $\mathcal{R}_R$ (rationality) and $\mathcal{R}_I$ (invariance).*

Since there might exist multiple optimal solutions to (P1), a tie breaking rule might be needed to decide on the method's output, Ψ_{MER}. We account for this fact in the experiments from Section 5. Note that the constraints in (P1) are quite restrictive, leading to blame assignments that typically distribute very little blame in total. The amount of total blame assigned is important for explanatory power. Namely, a trivial blame attribution method that assigns the score of 0 to every agent satisfies all of the properties from the previous section except $\mathcal{R}_E$ (efficiency), but provides no information regarding the agents' contributions to the outcome.

3.2 Marginal Contribution

Another intuitive blame assignment method is what we refer to as *marginal contribution*. This method simply quantifies an agent's potential to increase the performance of the system, assuming that the other agents keep their policies fixed. That is, the blame assigned to agent i is equal to $\beta_i = \Delta_i$. The following properties hold:

Proposition 2. *$\Psi_{MC}(M, \pi^b) = (\Delta_1, ..., \Delta_n)$ satisfies $\mathcal{R}_S$ (symmetry), $\mathcal{R}_I$ (invariance), $\mathcal{R}_{CM}$ (contribution monotonicity) and $\mathcal{R}_{PerM}$ (performance monotonicity).*

Unlike max-efficient rationality, marginal contribution does not satisfy validity, i.e., it can over-blame a group of agents by assigning them total score that exceeds the improvement they can achieve, i.e., Δ. Given that an agent's marginal inefficiency is not always a good indicator of the agent's influence on the system's performance, this method can be highly inefficient (distributing very little blame) when coordination among agents is required, as we show in Section 5.

3.3 Shapley Value and Banzhaf Index

In the context of the sequential decision making setting studied in this paper, Shapley value can be defined as $\beta = \Psi_{SV}(M, \pi^b)$ such that

$$\beta_i = \sum_{S \subseteq \{1,...,n\} \setminus \{i\}} w_S \cdot \left[J(\pi^{*|\pi^b}_{S \cup \{i\}}, {\pi^b}_{-S \cup \{i\}}) - J(\pi^{*|\pi^b}_{S}, {\pi^b}_{-S}) \right], \tag{1}$$

where coefficients w_S are set to $w_S = \frac{|S|!(n-|S|-1)!}{n!}$. We restate (and in Appendix H, prove the claim for our setting) a well known uniqueness result for Shapley value:

Theorem 1. *[51] $\Psi_{SV}(M, \pi^b) = (\beta_1, ..., \beta_n)$, where β_i is defined by Eq. (1) and $w_S = \frac{|S|!(n-|S|-1)!}{n!}$, is a unique blame attribution method satisfying $\mathcal{R}_E$ (efficiency), $\mathcal{R}_S$ (symmetry) and $\mathcal{R}_{CM}$ (contribution monotonicity). Additionally, Ψ_{SV} satisfies $\mathcal{R}_V$ (validity) and $\mathcal{R}_I$ (invariance).*

As we show in Section 5, Shapley value does not satisfy properties $\mathcal{R}_R$ (rationality) nor $\mathcal{R}_{PerM}$ (performance monotonicity).

Banzhaf index, denoted by Ψ_{BI}, is similar to Shapley value, and in fact, it has the same functional form but different coefficients ($w_S = \frac{1}{2^{n-1}}$), leading to a slightly different uniqueness result. Appendix C discusses Banzhaf index and its properties in greater detail. Here, we note that Banzhaf index is equivalent to Shapley value for two agents. However, in general, Banzhaf index does not satisfy $\mathcal{R}_E$ (efficiency), but a version of it, called 2-efficiency [52]. As it is the case with Shapley value, Banzhaf index does not satisfy $\mathcal{R}_{PerM}$ (performance monotonicity) nor $\mathcal{R}_R$ (rationality). Interestingly, $\mathcal{R}_V$ (validity) might also not hold (see Section 5).

3.4 Average Participation

Motivated by the fact that $\mathcal{R}_{PerM}$ (performance monotonicity) is important for incentivizing good performance and $\mathcal{R}_E$ (efficiency) is important for explanatory power, we introduce a novel blame assignment method, which can be seen as a combination of marginal contribution and Shapley value. We first show the following result, which shows that there is an inherent trade-off between $\mathcal{R}_{PerM}$ and $\mathcal{R}_E$, assuming $\mathcal{R}_S$ (symmetry) and $\mathcal{R}_I$ (invariance) hold.

Proposition 3. *No blame attribution method Ψ satisfies $\mathcal{R}_E$ (efficiency), $\mathcal{R}_S$ (symmetry), $\mathcal{R}_I$ (invariance) and $\mathcal{R}_{PerM}$ (performance monotonicity).*

Given this result, we instead consider two new properties $\mathcal{R}_{AE}$ (average efficiency) and $\mathcal{R}_{cPerM}$ (c-performance monotonicity), which are weaker variants of $\mathcal{R}_E$ (efficiency) and $\mathcal{R}_{PerM}$ (performance

monotonicity) respectively. Importantly, $\mathcal{R}_{AE}$ is not satisfied by Ψ_{MC} and $\mathcal{R}_{cPerM}$ by Ψ_{SV}. In addition, we also consider two variants of $\mathcal{R}_{CM}$ (contribution monotonicity): $\mathcal{R}_{cParM}$ (c-participation monotonicity) and $\mathcal{R}_{RcParM}$ (relative c-participation monotonicity). To define the new properties, we introduce a contribution function $c : \mathcal{M} \times \Pi \times \{1, ..., n\} \to \{0, 1\}$ that indicates whether an agent is *pivotal*, i.e., marginally contributes to the inefficiency of some subset of $\{1, ..., n\}$:

$$c(M, \pi^b, i) = \begin{cases} 0 & \text{if } J(\pi^{*|\pi^b}_{S\cup\{i\}}, {\pi^b}_{-S\cup\{i\}}) = J(\pi^{*|\pi^b}_{S}, {\pi^b}_{-S}) \quad \forall S \subseteq \{1, ..., n\} \\ 1 & \text{otherwise} \end{cases}.$$

Alternatively, an agent i is pivotal if and only if its Shapley value is strictly greater than 0, i.e., $c(M, \pi^b, i) = \mathbb{1}[\beta_i > 0]$, where $\beta = \Psi_{SV}(M, \pi^b)$ and $\mathbb{1}[.]$ is an indicator function. The new properties are then defined as follows:

- *Average efficiency* ($\mathcal{R}_{AE}$): Ψ satisfies $\mathcal{R}_{AE}$ (resp. ϵ-$\mathcal{R}_{AE}$) if for every M and π^b, $\sum_{i=1}^{n} \beta_i = \sum_{S\subseteq\{1,...,n\}} \frac{1}{2^n-1} \cdot \Delta_S$ (resp. $|\sum_{i=1}^{n} \beta_i - \sum_{S\subseteq\{1,...,n\}} \frac{1}{2^n-1} \cdot \Delta_S| \leq \epsilon$), where $\beta = \Psi(M, \pi^b)$.
- *c-Performance monotonicity* ($\mathcal{R}_{cPerM}$): Consider any MMDP M, and any ${\pi^b}_{-i}$, π_i and π'_i s.t. $J(\pi_i, {\pi^b}_{-i}) \leq J(\pi'_i, {\pi^b}_{-i})$ and $c(M, (\pi_i, {\pi^b}_{-i}), j) = c(M, (\pi'_i, {\pi^b}_{-i}), j)$ for every j. We say that Ψ satisfies $\mathcal{R}_{cPerM}$ (resp. ϵ-$\mathcal{R}_{cPerM}$) if $\beta_i \geq \beta'_i$ (resp. $\beta_i \geq \beta'_i - \epsilon$) where $\beta = \Psi(M, (\pi_i, {\pi^b}_{-i}))$ and $\beta' = \Psi(M, (\pi'_i, {\pi^b}_{-i}))$.
- *c-Participation monotonicity* ($\mathcal{R}_{cParM}$): Ψ satisfies $\mathcal{R}_{cParM}$ (resp. ϵ-$\mathcal{R}_{cParM}$) if for every $(M^1, {\pi^b}^1)$ and $(M^2, {\pi^b}^2)$ s.t. $c(M^1, {\pi^b}^1, i) = c(M^2, {\pi^b}^2, i)$ for every i, ${\beta_j}^1 \geq {\beta_j}^2$ (resp. ${\beta_j}^1 \geq {\beta_j}^2 - \epsilon$) whenever $\Delta^1_{S\cup\{j\}} \geq \Delta^2_{S\cup\{j\}}$ for all S, where $\beta^1 = \Psi(M^1, {\pi^b}^1)$ and $\beta^2 = \Psi(M^2, {\pi^b}^2)$.
- *Relative c-participation monotonicity* ($\mathcal{R}_{RcParM}$): Ψ satisfies $\mathcal{R}_{RcParM}$ (resp. ϵ-$\mathcal{R}_{RcParM}$) if for every $(M^1, {\pi^b}^1)$ and $(M^2, {\pi^b}^2)$ s.t. $c(M^1, {\pi^b}^1, i) = c(M^2, {\pi^b}^2, i)$ for every i, ${\beta_j}^1 - {\beta_j}^2 \geq {\beta_k}^1 - {\beta_k}^2$ (resp. ${\beta_j}^1 - {\beta_j}^2 \geq {\beta_k}^1 - {\beta_k}^2 - \epsilon$) whenever $c(M^1, {\pi^b}^1, j) = c(M^1, {\pi^b}^1, k)$ and $\Delta^1_{S\cup\{j\}} - \Delta^2_{S\cup\{j\}} \geq \Delta^1_{S\cup\{k\}} - \Delta^2_{S\cup\{k\}}$ for all $S \in \{1, ..., n\}\backslash\{j, k\}$, where $\beta^1 = \Psi(M^1, {\pi^b}^1)$ and $\beta^2 = \Psi(M^2, {\pi^b}^2)$.

Before describing the main results of this subsection, we briefly outline the intuition behind the above definitions. $\mathcal{R}_{AE}$ (average efficiency) is similar to $\mathcal{R}_E$ (efficiency), however it requires less total blame to be distributed. Whereas $\mathcal{R}_E$ requires that the total blame is equal to the total inefficiency Δ, $\mathcal{R}_{AE}$ requires that the total blame is equal to the average marginal inefficiency of subsets of agents, i.e., the average value of Δ_S.[6] Compared to $\mathcal{R}_{PerM}$ (performance monotonicity), $\mathcal{R}_{cPerM}$ (c-performance monotonicity) additionally accounts for the pivotality of agents through contribution function c, treating each set of pivotal agents as a separate case. $\mathcal{R}_{cParM}$ (c-participation monotonicity) accounts for agents' pivotality in a similar manner. Moreover, $\mathcal{R}_{cParM}$ resembles contribution monotonicity $\mathcal{R}_{CM}$, but instead of requiring blame monotonicity to hold w.r.t. the agent's influence on the marginal inefficiency of subsets S (i.e., $\Delta_{S\cup\{i\}} - \Delta_S$), it considers blame monotonicity w.r.t. the marginal inefficiency of subsets that contain the agent (i.e., $\Delta_{S\cup\{j\}}$). Relative c-participation monotonicity $\mathcal{R}_{RcParM}$ is similar to $\mathcal{R}_{cParM}$, but its blame monotonicity requirement is based on a pairwise comparison of agents with the same pivotality degree. In particular, $\mathcal{R}_{RcParM}$ requires that the blame increase is higher for an agent who is in subsets with a greater marginal inefficiency increase (i.e., ${\beta_j}^1 - {\beta_j}^2 \geq {\beta_k}^1 - {\beta_k}^2$ whenever $\Delta^1_{S\cup\{j\}} - \Delta^2_{S\cup\{j\}} \geq \Delta^1_{S\cup\{k\}} - \Delta^2_{S\cup\{k\}}$).

Average participation: Now, we describe the new blame assignment method, which we call *average participation*. This blame assignment method can be defined as $\beta = \Psi_{AP}(M, \pi^b)$ such that

$$\beta_i = \sum_{S\subseteq\{1,...,n\}\backslash\{i\}} w \cdot \frac{c(M, \pi^b, i)}{\sum_{j\in S} c(M, \pi^b, j) + 1} \cdot \Delta_{S\cup\{i\}}, \tag{2}$$

where coefficient w is set to $w = \frac{1}{2^n-1}$. Intuitively, Ψ_{AP} equally distributes blame for the marginal inefficiency of a subset of agents among the pivotal agents in that subset. Hence, an agent i that is pivotal receives blame for each subset $S \cup \{i\}$ equal to $\Delta_{S\cup\{i\}}$ divided by the number of pivotal agents in $S \cup \{i\}$ and scaled by coefficient w. Agents that are not pivotal, obtain 0 blame. Average participation uniquely satisfies the following properties.

[6] This average does not include $\Delta_\emptyset$, which is equal to 0. Note also that $\Delta \geq \Delta_S$ for every $S \subseteq \{1, ..., n\}$, so this average is upper bounded by Δ.

Theorem 2. *$\Psi_{AP}(M, \pi^b) = (\beta_1, ..., \beta_n)$, where β_i is defined by Eq. (2) and $w = \frac{1}{2^n - 1}$, is a unique blame attribution method that satisfies $\mathcal{R}_{AE}$ (average-efficiency), $\mathcal{R}_S$ (symmetry), $\mathcal{R}_I$ (invariance), $\mathcal{R}_{cParM}$ (c-participation monotonicity) and $\mathcal{R}_{RcParM}$ (relative c-participation monotonicity). Furthermore, Ψ_{AP} satisfies $\mathcal{R}_{cPerM}$ (c-performance monotonicity) and $\mathcal{R}_V$ (validity).*

Unlike marginal contribution, average participation is valid (never over-blames agents), however it satisfies a weaker version of performance monotonicity. Still, this version is not satisfied by Shapley value. On the other hand, Shapley value is efficient, unlike average participation, which satisfies a weaker requirement—average efficiency. We also showcase these trade-offs in Section 5.

4 Blame Attribution under Uncertainty

In this section, we study blame attribution methods that do not have direct access to π^b. As mentioned in Section 2, we focus on the case where the knowledge about π^b is defined by the uncertainty set $\mathcal{P}(\pi^b)$, and it is defined state-wise so that each state is associated with a set of probability measures on $\mathcal{A}$ identifying plausible candidates for $\pi^b(\cdot|s)$.[7] We denote $\pi \in \mathcal{P}(\pi^b)$ if π is plausible by $\mathcal{P}(\pi^b)$.

4.1 Shapley Value under Uncertainty

In explaining approaches to handling uncertainty, we focus on Shapley value. Arguably, the simplest way to operate under uncertainty is to derive a point estimate of π^b, denoted by $\widehat{\pi}^b$,[8] and apply Ψ_{SV} on this estimate to obtain blame assignment $\widehat{\beta} = \Psi_{SV}(M, \widehat{\pi}^b)$. Albeit being simple, this approach does not satisfy desirable properties, most notably, $\mathcal{R}_V$ (validity) and $\mathcal{R}_{BC}$ (Blackstone consistency).

Validity. Now, note that $\widehat{\beta} = \Psi_{SV}(M, \widehat{\pi}^b)$ satisfies $\sum_{i=1}^n \widehat{\beta}_i = J(\pi^*) - J(\widehat{\pi}^b)$. Therefore, instead of relying on a point estimate $\widehat{\pi}^b$, we could utilize a policy $\widehat{\pi}^b$ for which $J(\pi^*) - J(\widehat{\pi}^b) \leq \Delta$. Namely, in that case $\widehat{\beta} = \Psi_{SV}(M, \widehat{\pi}^b)$ results in a blame assignment that satisfies $\mathcal{R}_V$ (validity). Since this inequality holds for a solution to the optimization problem $\max_{\pi \in \mathcal{P}(\pi^b)} J(\pi)$, we obtain:

Proposition 4. *Let $\widehat{\pi}^b$ be a solution to the optimization problem $\max_{\pi \in \mathcal{P}(\pi^b)} J(\pi)$. Then $\widehat{\Psi}_{SV,V}(M, \mathcal{P}(\pi^b)) = \Psi_{SV}(M, \widehat{\pi}^b)$ satisfies $\mathcal{R}_V$ (validity).*

Blackstone consistency. As we show in Section 5, although $\widehat{\Psi}_{SV,V}$ is valid, it might not be Blackstone consistent w.r.t. Ψ_{SV}. In particular, although the total blame is never overestimated, an agent i might receive higher blame than it would receive under Ψ_{SV}. To ensure Blackstone consistency, we can assign blame to agent i equal to $\min_{\pi \in \mathcal{P}(\pi^b)} {\beta_i}^\pi$ s.t. $\beta^\pi = \Psi_{SV}(M, \pi)$. Together with Eq. (1), this implies that agent i's blame is obtained by solving

$$\min_{\pi \in \mathcal{P}(\pi^b)} \sum_{S \subseteq \{1,...,n\} \setminus \{i\}} w_S \cdot \left[J(\pi^{*|\pi}_{S \cup \{i\}}, \pi_{-S \cup \{i\}}) - J(\pi^{*|\pi}_S, \pi_{-S}) \right], \tag{P2}$$

where $w_S = \frac{|S|!(n-|S|-1)!}{n!}$ and $\pi^{*|\pi}_S \in \arg\max_{\pi'_S} J(\pi'_S, \pi_{-S})$. We have the following result:

Proposition 5. *Let β^i_i be the minimum value of the objective in* (P2)*. Then $\widehat{\Psi}_{SV,BC}(M, \mathcal{P}(\pi^b)) = (\beta^1_1, ..., \beta^n_n)$ satisfies $\mathcal{R}_V$ (validity) and $\mathcal{R}_{BC}(\Psi_{SV})$ (Blackstone consistency $w.r.t.$ $\Psi_{SV}(M, \pi^b)$).*

Note that $\widehat{\Psi}_{SV,BC}$ distributes less total blame than $\widehat{\Psi}_{SV,V}$, since it takes the worst case perspective for each agent separately, while under $\widehat{\Psi}_{SV,V}$ the blame assigned to all agents is computed with the same joint behavior policy. Moreover, the objective function in (P2) is more complex than in classical robust MDP settings [47, 48], making classical approaches for robust MDPs hard to apply. In practice, we can relax (P2) and optimize a lower bound of the objective; this preserves $\mathcal{R}_{BC}(\Psi_{SV})$, but at the expense of distributing less blame to the agents. In our experiments, we solve $\min_{\pi \in \mathcal{P}'(\pi^b)} J(\pi^{*|\pi}_{S \cup \{i\}}, \pi_{-S \cup \{i\}})$ and $\max_{\pi \in \mathcal{P}'(\pi^b)} J(\pi^{*|\pi}_S, \pi_{-S})$ for each subset S and with appropriately chosen $\mathcal{P}'(\pi^b) \supseteq \mathcal{P}(\pi^b)$ (see Appendix D), and we apply Eq. (1) to obtain the blame

[7] Such definition implies a rectangularity of the uncertainty set [47, 49].

[8] For example, this estimate can be derived from data containing the agents' trajectories.

assignment. This implies that agent i's blame is obtained by solving

$$\sum_{S\subseteq\{1,\dots,n\}\setminus\{i\}} w_S \cdot \left[\min_{\pi\in\mathcal{P}'(\pi^b)} J(\pi^{*|\pi}_{S\cup\{i\}}, \pi_{-S\cup\{i\}}) - \max_{\pi\in\mathcal{P}'(\pi^b)} J(\pi^{*|\pi}_{S}, \pi_{-S})\right].$$

Other Blame Attribution Methods. Similar approaches also work for other blame assignment methods discussed in Section 3. For example, and focusing on Blackstone consistency, $\widehat{\Psi}_{BI,BC}(M, \mathcal{P}(\pi^b))$ can be obtained in the same way as $\widehat{\Psi}_{SV,BC}(M, \mathcal{P}(\pi^b))$, but with $w_S = \frac{1}{2^{n-1}}$, while $\widehat{\Psi}_{MC,BC}(M, \mathcal{P}(\pi^b))$ can be implemented as $\widehat{\Psi}_{MC,BC}(M, \mathcal{P}(\pi^b)) = (\tilde{\Delta}_1, ..., \tilde{\Delta}_n)$ where $\tilde{\Delta}_i = \min_{\pi\in\mathcal{P}'(\pi^b)} J(\pi_i^{*|\pi}, \pi_{-i}) - \max_{\pi\in\mathcal{P}'(\pi^b)} J(\pi)$. Implementing Blackstone consistent $\widehat{\Psi}_{MER,BC}(M, \mathcal{P}(\pi^b))$ and $\widehat{\Psi}_{AP,BC}(M, \mathcal{P}(\pi^b))$ is more nuanced, and we discuss it in Appendix D.

4.2 Characterization Result

Notice that the described Blackstone consistent methods $\widehat{\Psi}(M, \mathcal{P}(\pi^b))$ are not guaranteed to satisfy the properties that their counterparts $\Psi(M, \pi^b)$ satisfy. However, as long as $\widehat{\Psi}(M, \mathcal{P}(\pi^b))$ and $\Psi(M, \pi^b)$ output similar enough blame assignments, properties that hold under $\Psi(M, \pi^b)$ will approximately hold under $\widehat{\Psi}(M, \mathcal{P}(\pi^b))$. More formally, we have the following results.

Theorem 3. *Consider $\widehat{\Psi}$ and Ψ s.t. $\left\|\widehat{\Psi}(M, \mathcal{P}(\pi^b)) - \Psi(M, \pi^b)\right\|_1 \leq \epsilon$ for any M, π^b, and $\mathcal{P}(\pi^b)$. Then if Ψ satisfies a property $\mathcal{R} \in \{\mathcal{R}_V, \mathcal{R}_E, \mathcal{R}_R, \mathcal{R}_S, \mathcal{R}_I, \mathcal{R}_{AE}\}$, $\widehat{\Psi}$ satisfies ϵ-$\mathcal{R}$. Moreover, if Ψ satisfies a property $\mathcal{R} \in \{\mathcal{R}_{CM}, \mathcal{R}_{PerM}, \mathcal{R}_{cPerM}, \mathcal{R}_{cParM}, \mathcal{R}_{RcParM}\}$, $\widehat{\Psi}$ satisfies 2ϵ-$\mathcal{R}$.*

This theorem allows us to quantify the robustness of the blame attribution methods—the closer $\widehat{\Psi}$ is to Ψ, the more robust it is to uncertainty. Interestingly, a trivial blame attribution method that assigns 0 blame to all the agents is robust in this sense. However, as we already mentioned, this trivial blame assignment is not informative as it does not attribute any blame. In fact, if agents receive no penalties for bad behavior, such a blame attribution method might have adverse effects. We provide a broader discussion on the negative side-effects of under-blaming in Appendix F. Importantly, this example suggests that efficiency (in a broad sense, i.e., how much blame is being distributed) and robustness are sometimes at odds, which we also demonstrate in the experiments.

5 Experiments

To demonstrate the efficacy of the studied blame attribution methods, we consider two environments, *Gridworld* and *Graph*, depicted in Fig. 1 and Fig. 2. Both environments are adapted from [53] and modified to be multi-agent. The experiments evaluate blame attribution methods along three axis:

- *Performance monotonicity:* First, we test blame attribution methods for the $\mathcal{R}_{PerM}$ (performance monotonicity) property, which we deem important for accountability. To do that, we consider the gridworld environment: this is a two-agent environment in which one of the agents, A_2, optimizes its policy using a model of the other agent, A_1. Importantly, by controlling the correctness of A_2's model of A_1, we can validate whether a blame attribution method satisfies $\mathcal{R}_{PerM}$. Namely, if A_2 does not receive the minimum blame when its model of A_1 is the correct model, the corresponding method is not performance incentivizing, i.e., it does not satisfy $\mathcal{R}_{PerM}$.
- *Coordination:* Second, we evaluate the efficacy of blame attribution methods when a higher degree of coordination among agents is needed to yield improvements over the baseline behavior. For this, we consider the graph environment, which includes configurations where an agent cannot improve the joint performance by unilaterally changing its policy. Thus, this environment is suitable for evaluating whether blame attribution methods incorporate more nuanced counterfactual reasoning.
- *Robustness:* Finally, we evaluate the robustness of blame attribution methods under uncertainty. In this case, both environments (Gridworld and Graph) are used for testing purposes, and we control for the level of uncertainty over the agents' behavior policies.

Appendix E provides more details on the experimental setup and implementation. Below we provide a more detailed description of the considered environments and discuss our findings.

Environment 1: This is a gridworld environment, in which two agents control the same actor but with different priorities. In the single-agent version of the environment, an agent, agent A_1, controls the movement of the actor. In our multi-agent version, there is an additional agent, agent A_2, who can intervene and override A_1's actions. The two agents select their actions simultaneously. Cells denoted with S are the initial states, blank cells indicate areas of small negative reward, F cells indicate areas of slightly increased cost and H cells are areas of severe penalty. The cell denoted by G is the terminal state of the environment and has a positive reward. When agent A_2 intervenes in some state, the actor takes the action that an optimal policy would select in the single-agent mode, but also pays a cost of intervention C. The behavior policy ${\pi^b}_1$ of agent A_1 is parameterized by variable α, which specifies the probability that A_1 takes an action determined by an optimal single-agent policy, instead of its personal policy. The personal policy of A_1 is a mixture of an optimal single-agent policy for correctly specified costs and a single-agent policy that is optimal but for misspecified costs of F and H cells—it assumes that they have the same cost as the blank cells. A_2's behavior policy ${\pi^b}_2$ optimizes the expected discounted return and is trained with a model of A_1 specified by the true personal policy of A_1 and variable α' (not necessarily equal to α). A_2 is meant to rectify potential mistakes of A_1 that could inflict cost greater than C. In $\mathcal{R}_{PerM}$ experiments we set $\alpha = 0.4$. In robustness experiments, we only consider uncertainty over the personal policy of A_1, and we set $\alpha = 0.2$ and $\alpha' = 0.5$.

S	S	S	S	S	S	S	S	S
S		F		H				
S			H			F		
S	F				H		F	
S			H			F		
S	H	H		F		H		
S	H			H		H		
S			H		F		G	

Fig. 1. Gridworld

Performance monotonicity: Fig. 3a validates our theoretical results regarding $\mathcal{R}_{PerM}$ (performance monotonicity). More specifically, methods Ψ_{AP} and Ψ_{MC} assign the minimum blame to A_2 when it acts optimally w.r.t. the true policy of A_1 , i.e., when $\alpha' = \alpha$. However, this is not the case for methods Ψ_{SV} and Ψ_{BI}, which implies that these methods are not incentivizing A_2 to act optimally w.r.t. its belief about A_1. Ψ_{MER} and Ψ_{BI} assign the same blame to A_2 as Ψ_{MC} and Ψ_{SV}, respectively.

Environment 2: This is a graph environment in which 4 agents simultaneously select actions. The graph consists of one starting and one terminal node, as well as 8 intermediate nodes that can be grouped according to their index number; nodes with even index number are located on the upper level of the graph and nodes with odd index number on the lower level. At each time-step every agent chooses to take either action 0 and move to the upper level or action 1 and move to the lower level. We test multiple variants of this environment, each of which defines a different reward function. In all variants, the reward at each time-step is $+1$ if some formation constraint is satisfied and -1 if not. In the first set of experiments (Coordination), we consider 4 different formation constraints: in formation constraint $m \in \{1, ..., 4\}$, at least m agents need to select action 1 for the constraint to be satisfied. Each behavior policy ${\pi^b}_i$ takes action $a_i = 0$ in every node. In the second set of experiments (Robustness), we consider one formation constraint that is satisfied if the agents are arranged equally between the two levels. In states where agents are balanced between the levels, each behavior policy ${\pi^b}_i$ takes the action from the previous time-step with probability p_i; in unbalanced states, the action that leads to the level with the least number of agents is taken with probability p_i.

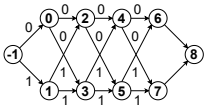
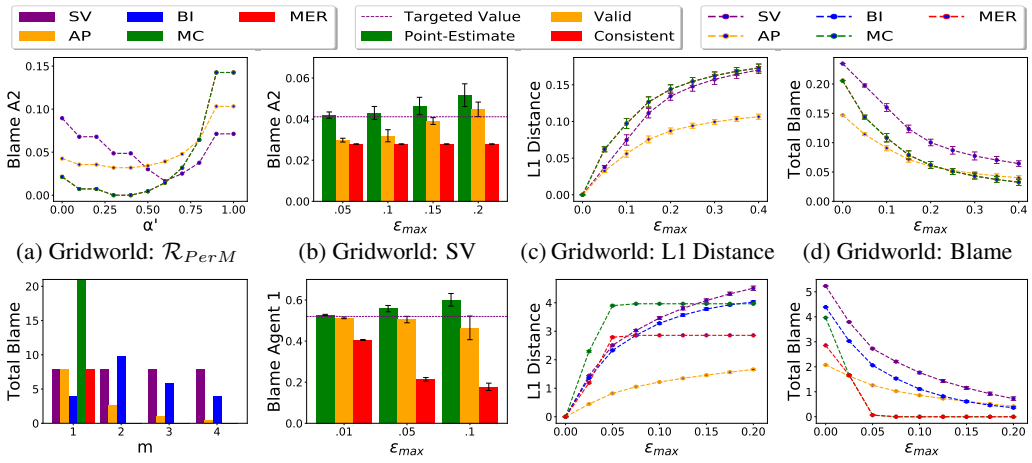


Fig. 2. Graph

Coordination: Fig. 3e shows how much blame in total the blame attribution methods assign for the four different levels of required coordination ($m = 1, ..., 4$). Observe, that when the constraint can be satisfied by every agent ($m = 1$), Ψ_{MC} violates $\mathcal{R}_V$ (validity). For $m = 2$, Ψ_{MER} and Ψ_{MC} assign zero blame to all agents, while Ψ_{BI} violates $\mathcal{R}_V$ (validity). Although always valid, Ψ_{AP} assigns significantly less blame as m increases. Ψ_{SV} is always efficient, and its total blame does not vary with m. Ψ_{SV}, Ψ_{BI}, Ψ_{MC} and Ψ_{AP} do not satisfy $\mathcal{R}_R$ (they assign more total blame than Ψ_{MER}).

Robustness: We test the robustness of the blame attribution methods by controlling the amount of uncertainty in the estimates of the agents' behavior policies. To model uncertainty, we consider maximum estimation error ϵ_{max}, and to obtain uncertainty sets $\mathcal{P}(\pi^b)$, we sample (uniformly at random) $\widehat{\pi}^b_i(s)$ such that $\frac{1}{2}\left\|\widehat{\pi}^b_i(s) - {\pi^b}_i(s)\right\|_1 \leq \epsilon_{max}$. Moreover, $\mathcal{P}(\pi^b, s)$ contains all policies π such that $\frac{1}{2}\left\|\widehat{\pi}^b_i(s) - \pi_i(s)\right\|_1 \leq \epsilon_{max}$. In our experiments, we take $\widehat{\pi}^b$ to be the point estimate of π^b.

Comparing estimation approaches: Fig. 3b and 3f show how the approaches for estimating $\widehat{\Psi}_{SV}$ from Section 4 fare under different levels of uncertainty. The point estimate approach typically over-blames an agent and the amount of over-blaming increases with the level of uncertainty. $\widehat{\Psi}_{SV,BC}$ never over-blames any agent, but it becomes less efficient (in distributing blame) as ϵ_{max} increases (Fig. 3f). $\widehat{\Psi}_{SV,V}$ is more efficient than $\widehat{\Psi}_{SV,BC}$, but violates $\mathcal{R}_{BC}$ (Blackstone consistency) (Fig. 3b).



(a) Gridworld: $\mathcal{R}_{PerM}$ (b) Gridworld: SV (c) Gridworld: L1 Distance (d) Gridworld: Blame

(e) Graph: Coordination (f) Graph: SV (g) Graph: L1 Distance (h) Graph: Blame

Fig. 3. Experimental results for the Gridworld and Graph environments. Plot (a) tests methods for $\mathcal{R}_{PerM}$. Plot (e) shows the effect of varying coordination level. Plots (b,c,d,f,g,h) show the effect of varying ϵ_{max} in different Shapley value approaches (b,f) and blame attribution methods (c,d,g,h).

Comparing attribution approaches: Fig. 3c and 3g show for each consistent blame attribution method $\widehat{\Psi}$ from Section 4 the L_1 distance between its output and the output of its counterpart Ψ ("targeted assignment"). Fig. 3d and 3h show the total blame assigned by these methods. $\widehat{\Psi}_{AP,BC}$ consistently outperforms the other methods in terms of the L_1 distance from its "targeted assignment". Compared to $\widehat{\Psi}_{AP,BC}$, $\widehat{\Psi}_{SV,BC}$ is consistently better in terms of efficiency (in distributing blame). Similar, albeit less prominent effects can be seen when comparing $\widehat{\Psi}_{AP,BC}$ and $\widehat{\Psi}_{BI,BC}$. These results indicate a tendency where efficiency (in distributing blame) and robustness are at odds, as we also discuss in Section 4.2. $\widehat{\Psi}_{MER,BC}$ and $\widehat{\Psi}_{MC,BC}$ assign zero total blame even for smaller ϵ_{max}, indicating that they are the least robust to uncertainty.

6 Conclusion

In summary, the focus of our work is to provide an overview of possible computational approaches for attributing blame in multi-agent sequential decision making. We discuss the strengths and weaknesses of different methods in order to guide practitioners and policy makers in designing tools that support accountability. We conclude that there is no single best choice for blame attribution methods, since there are inherent trade-offs among properties that one might consider important. Looking forward, we recognize several research directions that could address the limitations of our results, some of which we highlight here. a) In this work we primarily focused on the agents' joint return as the outcome of interest. However, it is often important to pinpoint actual causes that led to more fine grained outcomes. Utilizing a causal perspective would be beneficial in this regard and could link our results to prior work (e.g., [16]). b) We considered model-based approaches to blame assignment. Learning blame attribution directly from data (e.g., with model-free counterfactual RL) might be more practical in settings where an approximate model is hard to obtain. c) More generally, ensuring scalability both in the number of agents and the the richness of environments is one of the most important steps for making this work more widely applicable. We deem approaches from multi-agent RL as suitable candidate solutions for resolving this problem. d) We primarily studied blame assignment properties that are taken from or closely relate to those from the game theory literature. This list could be extended and include more principles from moral philosophy and law. For example, in this paper, we adopted a consequentialist approach to blame attribution, focusing on the outcomes of the agents' behavior. Alternatively, one could take a deontological perspective, and focus on the alignment of an agent's behavior with a set of rules. We further discuss different perspectives on blame attribution in Appendix F. Finally, we would like to draw particular attention to the fact that there is no universal prioritization of properties that applies to all blame attribution problems and hence treating any generic analysis like ours as panacea without further justification, might have a negative impact to the agents that are being blamed. To that end, we would like to emphasize that we see this work not as a final solution to the blame attribution problem, but as a starting point that shows challenges and trade-offs in distributing blame.

References

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2016.

[2] Amir E. Khandani, Adlar J. Kim, and Andrew W. Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010.

[3] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[4] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Kate Scott, Stuart Schieber, James Waldo, David Weinberger, et al. Accountability of AI under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*, 2017.

[5] Joshua A. Kroll, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. Accountable algorithms. *University of Pennsylvania Law Review*, 165:633, 2016.

[6] Maranke Wieringa. What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In *Conference on Fairness, Accountability, and Transparency*, pages 1–18, 2020.

[7] European Commission. Ethics Guidelines for Trustworthy Artificial Intelligence. URL: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai, 2019. [Online; accessed 15-January-2021].

[8] Mark Bovens. Analysing and assessing accountability: A conceptual framework. *European Law Journal*, 13(4):447–468, 2007.

[9] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016.

[10] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *IEEE Symposium on Security and Privacy*, pages 598–617, 2016.

[11] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *International Conference on Neural Information Processing Systems*, pages 4768–4777, 2017.

[12] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[13] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[14] Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pages 1507–1516, 2019.

[15] Hana Chockler and Joseph Y. Halpern. Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, 22:93–115, 2004.

[16] Joseph Y. Halpern. *Actual causality*. MiT Press, 2016.

[17] Joseph Y. Halpern and Max Kleiman-Weiner. Towards formal definitions of blameworthiness, intention, and moral responsibility. In *AAAI Conference on Artificial Intelligence*, page 1853–1860, 2018.

[18] Meir Friedenberg and Joseph Y. Halpern. Blameworthiness in multi-agent settings. In *AAAI Conference on Artificial Intelligence*, pages 525–532, 2019.

[19] Craig Boutilier. Planning, learning and coordination in multiagent decision processes. In *Conference on Theoretical Aspects of Rationality and Knowledge*, pages 195–210, 1996.

[20] John Von Neumann and Oskar Morgenstern. *Theory of games and economic behavior (commemorative edition)*. Princeton University Press, 2007.

[21] Kamal Jain and Mohammad Mahdian. Cost sharing. *Algorithmic Game Theory*, 15:385–410, 2007.

[22] Maria-Florina Balcan, Ariel D. Procaccia, and Yair Zick. Learning cooperative games. In *International Conference on Artificial Intelligence*, pages 475–481, 2015.

[23] Eric Balkanski, Umar Syed, and Sergei Vassilvitskii. Statistical cost sharing. In *International Conference on Neural Information Processing Systems*, pages 6222–6231, 2017.

[24] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J. Spanos. Towards efficient data valuation based on the Shapley value. In *International Conference on Artificial Intelligence and Statistics*, pages 1167–1176, 2019.

[25] Anish Agarwal, Munther Dahleh, and Tuhin Sarkar. A marketplace for data: An algorithmic solution. In *ACM Conference on Economics and Computation*, pages 701–726, 2019.

[26] Yoav Shoham and Kevin Leyton-Brown. *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press, 2008.

[27] Georgios Chalkiadakis and Craig Boutilier. Bayesian reinforcement learning for coalition formation under uncertainty. In *International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 1090–1097, 2004.

[28] Donald B. Gillies. Solutions to general non-zero-sum games. *Contributions to the Theory of Games*, 4:47–85, 1959.

[29] Lloyd S. Shapley. *17. A value for n-person games*. Princeton University Press, 2016.

[30] Lloyd S. Shapley and Martin Shubik. A method for evaluating the distribution of power in a committee system. *The American Political Science Review*, 48(3):787–792, 1954.

[31] John F. Banzhaf III. Weighted voting doesn't work: A mathematical analysis. *Rutgers Law Review*, 19:317, 1964.

[32] John F. Banzhaf III. One man, 3.312 votes: a mathematical analysis of the electoral college. *Villanova Law Review*, 13:304, 1968.

[33] William Blackstone and George Sharswood. *Commentaries on the Laws of England. In Four Books*. JB Lippincott, 1893.

[34] Thomas M. Scanlon. *Moral dimensions*. Harvard University Press, 2009.

[35] David Shoemaker. Attributability, answerability, and accountability: Toward a wider theory of moral responsibility. *Ethics*, 121(3):602–632, 2011.

[36] Ibo Van de Poel, Lambèr Royakkers, and Sjoerd D. Zwart. *Moral responsibility and the problem of many hands*. Routledge, 2015.

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

[38] Steve Torrance. Ethics and consciousness in artificial agents. *AI & Society*, 22(4):495–521, 2008.

[39] Peter M. Asaro. Robots and responsibility from a legal perspective. *IEEE*, 4(14):20–24, 2007.

[40] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. Human perceptions on moral responsibility of AI: A case study in AI-assisted bail decision-making. In *CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2021.

[41] Christel Baier, Florian Funke, and Rupak Majumdar. A game-theoretic account of responsibility allocation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 1773–1779, 2021.

[42] Marvin Minsky. Steps toward artificial intelligence. *Institute of Radio Engineers*, 49(1):8–30, 1961.

[43] Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT Press, 2018.

[44] Kagan Tumer and Adrian Agogino. Distributed agent-based air traffic flow management. In *International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 1–8, 2007.

[45] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *AAAI Conference on Artificial Intelligence*, pages 2974–2982, 2018.

[46] Jianhong Wang, Yuan Zhang, Tae-Kyun Kim, and Yunjie Gu. Shapley Q-value: A local reward approach to solve global reward games. In *AAAI Conference on Artificial Intelligence*, pages 7285–7292, 2020.

[47] Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

[48] Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[49] Aviv Tamar, Shie Mannor, and Huan Xu. Scaling up robust MDPs using function approximation. In *International Conference on Machine Learning*, pages 181–189, 2014.

[50] Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia. *Handbook of computational social choice*. Cambridge University Press, 2016.

[51] Peyton H. Young. Monotonic solutions of cooperative games. *International Journal of Game Theory*, 14(2):65–72, 1985.

[52] Marcin Malawski. Equal treatment, symmetry and Banzhaf value axiomatizations. *International Journal of Game Theory*, 31(1):47–67, 2002.

[53] Cameron Voloshin, Hoang M. Le, Nan Jiang, and Yisong Yue. Empirical study of off-policy policy evaluation for reinforcement learning. *arXiv preprint arXiv:1911.06854*, 2019.

[54] Anupam Datta, Deepak Garg, Dilsun Kaynar, Divya Sharma, and Arunesh Sinha. Program actions as actual causes: A building block for accountability. In *IEEE Computer Security Foundations Symposium*, pages 261–275, 2015.

[55] Simon Blackburn. *The Oxford dictionary of philosophy*. Oxford University Press, 2005.

[56] Samuel Scheffler et al. *Consequentialism and its Critics*. Oxford University Press, 1988.

[57] Immanuel Kant. *Groundwork of the Metaphysic of Morals*. Routledge, 2020.

[58] Gertrude Elizabeth Margaret Anscombe. Modern moral philosophy. *Philosophy*, 33(124):1–19, 1958.

[59] Michael S. Moore. *Placing blame: A theory of the criminal law*. Oxford University Press, 2010.

[60] Miklós Pintér. Young's axiomatization of the Shapley value: a new proof. *Annals of Operations Research*, 235(1):665–673, 2015.

A List of Appendices

In this section we provide a brief description of the content provided in the appendices of the paper.

- Appendix B provides a table that summarizes the the results in Section 3.
- Appendix C provides additional details on Banzhaf index.
- Appendix D provides additional details on blame attribution under uncertainty.
- Appendix E provides additional details on experimental setup and implementation.
- Appendix F provides an extended discussion on different perspectives on blame attribution and the negative side-effects of under-blaming agents.
- Appendix G contains the proofs of the proposition from Section 3 (Proposition 1, Proposition 2, and Proposition 3).
- Appendix H contains the proof of Theorem 1 from Section 3.
- Appendix I contains the proof of Theorem 2 from Section 3.
- Appendix J contains the proofs of the formal results from Section 4 (Proposition 4, Proposition 5, and Theorem 3).

B Table of Methods and Properties

In this section we provide a table that summarizes the results of Section 3 and describes which blame attribution methods satisfy which properties. We use (✓) to denote that a method does not satisfy the exact property but a weaker version of it.

	Ψ_{MER}	Ψ_{MC}	Ψ_{SV}	Ψ_{BI}	Ψ_{AP}
$\mathcal{R}_V$	✓		✓		✓
$\mathcal{R}_E$			✓		(✓)
$\mathcal{R}_R$	✓				
$\mathcal{R}_S$		✓	✓	✓	✓
$\mathcal{R}_I$	✓	✓	✓	✓	✓
$\mathcal{R}_{CM}$		✓	✓	✓	
$\mathcal{R}_{PerM}$		✓			(✓)

Table 1: Summary of the characterization results from Section 3

Method Ψ_{AP} satisfies properties $\mathcal{R}_{AE}$ and $\mathcal{R}_{cPerM}$ which are weaker versions of $\mathcal{R}_E$ and $\mathcal{R}_{PerM}$, respectively.

C Banzhaf Index

In this section, we discuss in a greater detail Banzhaf index and its properties. In the context of the sequential decision making setting studied in this paper, Banzhaf Index can be defined as $\beta = \Psi_{BI}(M, \pi^b)$ such that

$$\beta_i = \sum_{S \subseteq \{1,\dots,n\} \setminus \{i\}} w_S \cdot \left[J(\pi^{*|\pi^b}_{S \cup \{i\}}, {\pi^b}_{-S \cup \{i\}}) - J(\pi^{*|\pi^b}_{S}, {\pi^b}_{-S}) \right], \tag{3}$$

where coefficients w_S are set to $w_S = \frac{1}{2^{n-1}}$. The following properties hold:

Proposition 6. *$\Psi_{BI}(M, \pi^b) = (\beta_1, ..., \beta_n)$, where β_i is defined by Eq. (3) and $w_S = \frac{1}{2^{n-1}}$, is a blame attribution method satisfying $\mathcal{R}_S$ (symmetry), $\mathcal{R}_I$ (invariance) and $\mathcal{R}_{CM}$ (contribution monotonicity).*

Proof. First, notice that Banzhaf Index can be redefined as $\beta = \Psi_{BI}(M, \pi^b)$ such that:

$$\beta_i = \sum_{S \subseteq \{1,\dots,n\} \setminus \{i\}} w_S \cdot \left[\Delta_{S \cup \{i\}} - \Delta_S \right]. \tag{4}$$

We prove the properties as follows:

- $\mathcal{R}_S$ (symmetry): Consider M, π^b, and agents i and j such that $\Delta_{S\cup\{i\}} = \Delta_{S\cup\{j\}}$ for all $S \subseteq \{1,...,n\}\backslash\{i,j\}$. Notice that $\Delta_{S\cup\{i\}} - \Delta_S = \Delta_{S\cup\{j\}} - \Delta_S$ and $\Delta_{S\cup\{i,j\}} - \Delta_{S\cup\{j\}} = \Delta_{S\cup\{i,j\}} - \Delta_{S\cup\{i\}}$ for all $S \subseteq \{1,...,n\}\backslash\{i,j\}$. Given the definition of $\beta = \Psi_{BI}(M, \pi^b)$, this implies that $\beta_i = \beta_j$, and hence property $\mathcal{R}_S$ (symmetry) is satisfied.
- $\mathcal{R}_I$ (invariance): Consider M, π^b, and agent i such that $\Delta_{S\cup\{i\}} = \Delta_S$ for all S. Given the definition of $\beta = \Psi_{BI}(M, \pi^b)$, this implies that $\beta_i = 0$, and hence property $\mathcal{R}_I$ (invariance) is satisfied.
- $\mathcal{R}_{CM}$ (contribution monotonicity): Consider M^1, ${\pi^b}^1$, M^2, ${\pi^b}^2$, and agent i such that $\Delta^1_{S\cup\{i\}} - \Delta^1_S \geq \Delta^2_{S\cup\{i\}} - \Delta^2_S$ for all S. By using the definitions of $\beta^1 = \Psi_{BI}(M^1, {\pi^b}^1)$ and $\beta^2 = \Psi_{BI}(M^2, {\pi^b}^2)$, this implies that:

$$\begin{aligned}{\beta_i}^1 = &\sum_{S\subseteq\{1,...,n\}\backslash\{i\}} w_S \cdot \left[\Delta^1_{S\cup\{i\}} - \Delta^1_S\right] \geq \\ \geq &\sum_{S\subseteq\{1,...,n\}\backslash\{i\}} w_S \cdot \left[\Delta^2_{S\cup\{i\}} - \Delta^2_S\right] = \\ = &{\beta_i}^2,\end{aligned}$$

and hence property $\mathcal{R}_{CM}$ (contribution monotonicity) is satisfied. □

In general, Banzhaf index satisfies a property called 2-efficiency [52] which leads to a slightly different uniqueness result than the one of Theorem 1. This property and the corresponding analysis are out of the scope of this paper, and we refer the reader to [52, 54] for more details.

D Additional Information on Blame Attribution under Uncertainty

In this section, we provide additional information on the optimization problems defined in Section 4.1 and the implementation of Blackstone consistent $\widehat{\Psi}_{MER,BC}(M, \mathcal{P}(\pi^b))$ and $\widehat{\Psi}_{AP,BC}(M, \mathcal{P}(\pi^b))$.

D.1 Implementation of Optimization Problems

In this section, we provide implementation details on the optimization problems defined in Section 4.1, for obtaining Valid and Blackstone consistent blame attribution methods. More specifically, we focus on the optimization problems $\min_{\pi\in\mathcal{P}'(\pi^b)} J(\pi^{*|\pi}_{S\cup\{i\}}, \pi_{-S\cup\{i\}})$ and $\max_{\pi\in\mathcal{P}'(\pi^b)} J(\pi^{*|\pi}_S, \pi_{-S})$, where $S \subseteq \{1,...,n\}$ and $\mathcal{P}'(\pi^b) \supseteq \mathcal{P}(\pi^b)$. We consider

$$\mathcal{P}(\pi^b) = \Bigg\{\pi | \pi(a|s) = \pi_1(a_1|s)\cdots\pi_n(a_n|s), \frac{1}{2}\cdot\left\|\pi_i(\cdot|s) - \pi^{bas}_i(\cdot|s)\right\|_1 \leq C, 0 \leq \pi_i(a_i|s) \leq 1, \sum_{a_i\in\mathcal{A}_i}\pi_i(a_i|s) = 1\Bigg\},$$

where C is a non-negative constant and π^{bas} is a baseline joint policy. In specific cases, we can set $\mathcal{P}'(\pi^b) = \mathcal{P}(\pi^b)$ and we discuss these cases below. In general, to more directly relate the optimization problems to prior work on robust optimization in MDPs [47, 48], we relax the constraint that π factorizes to $\pi(a|s) = \pi_1(a_1|s)\cdots\pi_n(a_n|s)$, and consider

$$\mathcal{P}'(\pi^b) = \Bigg\{\pi | \prod_{i=1}^n \max(\pi^{bas}_i(a_i|s) - C, 0) \leq \pi(a_1,...,a_n|s) \leq \prod_{i=1}^n \min(\pi^{bas}_i(a_i|s) + C, 1), \sum_{(a_1,...,a_n)\in\mathcal{A}}\pi(a_1,...,a_n|s) = 1\Bigg\}.$$

Notice that since $\sum_{a_i\in\mathcal{A}_i}\pi^{bas}_i(a_i|s) = 1$, we have that $\pi^{bas}_i(a_i|s) - C \leq \pi_i(a_i|s) \leq \pi^{bas}_i(a_i|s) + C$ for every $\pi \in \mathcal{P}(\pi^b)$, and hence $\mathcal{P}''(\pi^b) \supseteq \mathcal{P}(\pi^b)$, where

$$\mathcal{P}''(\pi^b) = \Bigg\{\pi | \pi(a|s) = \pi_1(a_1|s)\cdots\pi_n(a_n|s), \max(\pi^{bas}_i(a_i|s) - C, 0) \leq \pi_i(a_i|s)$$

$$\leq \min(\pi_i^{bas}(a_i|s) + C, 1), \sum_{a_i \in \mathcal{A}_i} \pi_i(a_i|s) = 1\Big\}.$$

Importantly, $\mathcal{P}'(\pi^b) \supseteq \mathcal{P}''(\pi^b)$ implies that $\mathcal{P}'(\pi^b) \supseteq \mathcal{P}(\pi^b)$, which means that $\max_{\pi \in \mathcal{P}'(\pi^b)} J(\pi_S^{*|\pi}, \pi_{-S})$ upper bounds $\max_{\pi \in \mathcal{P}(\pi^b)} J(\pi_S^{*|\pi}, \pi_{-S})$ and $\min_{\pi \in \mathcal{P}'(\pi^b)} J(\pi_{S\cup\{i\}}^{*|\pi}, \pi_{-S\cup\{i\}})$ lower bounds $\min_{\pi \in \mathcal{P}(\pi^b)} J(\pi_{S\cup\{i\}}^{*|\pi}, \pi_{-S\cup\{i\}})$. Therefore, $\min_{\pi \in \mathcal{P}'(\pi^b)} J(\pi_{S\cup\{i\}}^{*|\pi}, \pi_{-S\cup\{i\}})$ and $\max_{\pi \in \mathcal{P}'(\pi^b)} J(\pi_S^{*|\pi}, \pi_{-S})$ can be used for deriving valid and Blackstone consistent blame assignments (e.g., by applying Eq. (1) with the obtained solutions). Next, we discuss how to solve these optimization problems.

While [47, 48] consider uncertainty over transitions dynamics instead of behavior policies, we can solve $\max_{\pi \in \mathcal{P}'(\pi^b)} J(\pi_S^{*|\pi}, \pi_{-S})$ and $\min_{\pi \in \mathcal{P}'(\pi^b)} J(\pi_{S\cup\{i\}}^{*|\pi}, \pi_{-S\cup\{i\}})$ by adapting their robust optimization techniques. To solve the optimization problem $\min_{\pi \in \mathcal{P}'(\pi^b)} J(\pi_{S\cup\{i\}}^{*|\pi}, \pi_{-S\cup\{i\}})$ for subset S, we apply the following recursion (in each iteration updating values for each state s):

$$\tilde{\pi}(\cdot|s) \leftarrow \operatorname*{arg\,min}_{\pi(\cdot|s) \in \mathcal{P}'(\pi^b, s)} \max_{a_{S\cup\{i\}}} \sum_{a_{-S\cup\{i\}}} \pi_{-S\cup\{i\}}(a_{-S\cup\{i\}}|s) \cdot \Big[R(s,a) + \gamma \cdot \sum_{s'} P(s,a,s') \cdot V^k(s')\Big],$$

$$V^{k+1}(s) \leftarrow \max_{a_{S\cup\{i\}}} \sum_{a_{-S\cup\{i\}}} \tilde{\pi}_{-S\cup\{i\}}(a_{-S\cup\{i\}}|s) \cdot \Big[R(s,a) + \gamma \cdot \sum_{s'} P(s,a,s') \cdot V^k(s')\Big],$$

for $k = 1, 2, \ldots$, where $V : \mathcal{S} \to \mathbb{R}_{\geq 0}$ is the value function, a_S denotes the joint action of agents S, a_{-S} denotes the joint action of agents $\{1, ..., n\} \backslash S$, and a is the joint action of all the agents. The optimization problem for finding $\tilde{\pi}$ can be solved via a linear program that minimizes a dummy variable which is constrained to be at least as large as

$$\sum_{a_{-S\cup\{i\}}} \pi_{-S\cup\{i\}}(a_{-S\cup\{i\}}|s) \cdot \Big[R(s,a) + \gamma \cdot \sum_{s'} P(s,a,s') \cdot V^k(s')\Big]$$

for all $a_{S\cup\{i\}}$. The optimization problem for finding V^{k+1} can be solved by simply searching over all possible $a_{S\cup\{i\}}$. Similarly, we can solve $\max_{\pi \in \mathcal{P}'(\pi^b)} J(\pi_S^{*|\pi}, \pi_{-S})$ with the following recursion:

$$\tilde{\pi}(\cdot|s) \leftarrow \operatorname*{arg\,max}_{\pi(\cdot|s) \in \mathcal{P}'(\pi^b, s)} \max_{a_S} \sum_{a_{-S}} \pi_{-S}(a_{-S}|s) \cdot \Big[R(s,a) + \gamma \cdot \sum_{s'} P(s,a,s') \cdot V^k(s')\Big],$$

$$V^{k+1}(s) \leftarrow \max_{a_S} \sum_{a_{-S}} \tilde{\pi}_{-S}(a_{-S}|s) \cdot \Big[R(s,a) + \gamma \cdot \sum_{s'} P(s,a,s') \cdot V^k(s')\Big],$$

for $k = 1, 2, \ldots$. The optimization problem for finding $\tilde{\pi}$ can be solved by searching over all a_S and selecting one that maximizes

$$\max_{\pi(\cdot|s) \in \mathcal{P}'(\pi^b, s)} \sum_{a_{-S}} \pi_{-S}(a_{-S}|s) \cdot \Big[R(s,a) + \gamma \cdot \sum_{s'} P(s,a,s') \cdot V^k(s')\Big]$$

—the solution to this problem gives us the corresponding $\tilde{\pi}$. The optimization problem for finding V^{k+1} can be solved by searching over all possible a_S. The two recursions described above define dynamic programming techniques that are analogs of those in [47, 48], but applied for uncertainty over behavior policies. They can be solved efficiently for smaller action spaces $\mathcal{A}$, e.g., as those in our experiments.

Now, in specific cases, we can set $\mathcal{P}'(\pi^b) = \mathcal{P}(\pi^b)$, which in turn can lead to more efficient blame assignments (since the estimates are tighter). We consider the following two cases:

- First, when there are only two agents in an MMDP, $-S \cup \{i\}$ contains at most one agent. Therefore, we could run the first recursion on $\{\pi_j | \frac{1}{2} \cdot \left\|\pi_j(\cdot|s) - \pi_j^{bas}(\cdot|s)\right\|_1 \leq C, 0 \leq \pi_j(a_j|s) \leq 1, \sum_{a_j \in \mathcal{A}_j} \pi_j(a_j|s) = 1\}$ instead of $\{\pi_j | \max(\pi_j^{bas}(a_j|s) - C, 0) \leq \pi_j(a_j|s) \leq \min(\pi_j^{bas}(a_j|s) + C, 1), \sum_{a_j \in \mathcal{A}_j} \pi_j(a_j|s) = 1\}$ and thus solve $\min_{\pi \in \mathcal{P}(\pi^b)} J(\pi_{S\cup\{i\}}^{*|\pi}, \pi_{-S\cup\{i\}})$.

Also in that case, $-S$ contains at most one agent whenever $S \neq \emptyset$, and hence we could run the second recursion on $\{\pi_j | \frac{1}{2} \cdot \left\|\pi_j(\cdot|s) - \pi_j^{bas}(\cdot|s)\right\|_1 \leq C, 0 \leq \pi_j(a_j|s) \leq 1, \sum_{a_j \in \mathcal{A}_j} \pi_j(a_j|s) = 1\}$, and thus solve $\max_{\pi \in \mathcal{P}(\pi^b)} J(\pi_S^{*|\pi}, \pi_{-S})$ for every $S \neq \emptyset$. In addition, when the optimal policies of one of the agents, agent i, are independent of which policy the other agent, agent j, follows we can directly compute an optimal policy for i on $\{\pi_i | \frac{1}{2} \cdot \left\|\pi_i(\cdot|s) - \pi_i^{bas}(\cdot|s)\right\|_1 \leq C, 0 \leq \pi_i(a_i|s) \leq 1, \sum_{a_i \in \mathcal{A}_i} \pi_i(a_i|s) = 1\}$, by fixing an arbitrary policy to agent j. Then, by fixing agent i to its optimal policy, we can can directly compute an optimal policy of agent j on $\{\pi_j | \frac{1}{2} \cdot \left\|\pi_j(\cdot|s) - \pi_j^{bas}(\cdot|s)\right\|_1 \leq C, 0 \leq \pi_j(a_j|s) \leq 1, \sum_{a_j \in \mathcal{A}_j} \pi_j(a_j|s) = 1\}$. This implies that we can run the second recursion directly on $\mathcal{P}(\pi^b)$ for $S = \emptyset$ and thus solve $\max_{\pi \in \mathcal{P}(\pi^b)} J(\pi)$. We use these facts in our experiments for the Gridworld environment, where the optimal policies of A_1 are independent of A_2's policy.

- Another specific case is when action spaces $\mathcal{A}_i$ are binary, and in this case, we can directly solve $\max_{\pi \in \mathcal{P}(\pi^b)} J(\pi_S^{*|\pi}, \pi_{-S})$. Namely, we can think of this optimization problem as searching for an optimal joint policy in an MMDP where the actions of agents $-S$ have reduced "influence". Since an optimal joint policy in the reduced MMDP is deterministic, the optimal solution to $\max_{\pi \in \mathcal{P}(\pi^b)} J(\pi_S^{*|\pi}, \pi_{-S})$ sets $\pi_j(a_j|s)$ of agent $j \in -S$ either to its maximum or its minimum value, $\pi_j^{bas}(a_j|s) + C$ and $\pi_j^{bas}(a_j|s) - C$ respectively. In the former case, this means that agent j chooses a_j in the MMDP with the reduced influence, in the latter, this means that agent j chooses the other action. We use this fact in our experiments for the Graph environment.

To conclude, in our experiments we directly solve the optimization problems $\min_{\pi \in \mathcal{P}(\pi^b)} J(\pi_{S \cup \{i\}}^{*|\pi}, \pi_{-S \cup \{i\}})$ and $\max_{\pi \in \mathcal{P}(\pi^b)} J(\pi_S^{*|\pi}, \pi_{-S})$ for the Gridworld environment, and $\max_{\pi \in \mathcal{P}(\pi^b)} J(\pi_S^{*|\pi}, \pi_{-S})$ for the Graph environment.

D.2 Max-Efficient Rationality and Average Participation under Uncertainty

In this section we discuss the implementation of Blackstone consistent $\widehat{\Psi}_{MER,BC}(M, \mathcal{P}(\pi^b))$ and $\widehat{\Psi}_{AP,BC}(M, \mathcal{P}(\pi^b))$ from Section 4.1. We begin with $\widehat{\Psi}_{MER,BC}(M, \mathcal{P}(\pi^b))$, which can be obtained by solving the optimization problem (P1) with Δ_S replaced by $\tilde{\Delta}_S = \min_{\pi \in \mathcal{P}'(\pi^b)} J(\pi_S^{*|\pi}, \pi_{-S}) - \max_{\pi \in \mathcal{P}'(\pi^b)} J(\pi)$. A solution to this optimization problem $\widehat{\beta}$ will for at least one solution β of (P1) (with Δ_S) satisfy $\widehat{\beta}_i \leq \beta_i$ for all i. In that sense, $\widehat{\Psi}_{MER,BC}(M, \mathcal{P}(\pi^b))$ satisfies $\mathcal{R}_{BC}(\Psi_{MER})$ (Blackstone consistency w.r.t. $\Psi_{MER}(M, \pi^b)$). However, note that $\mathcal{R}_{BC}(\Psi_{MER})$ might not hold if (P1) has multiple solutions (e.g., when calculating $\widehat{\Psi}_{MER,BC}$ or Ψ_{MER}) and we consider only one solution (e.g., obtained through a tie breaking rule).

Let us now consider $\widehat{\Psi}_{AP,BC}(M, \mathcal{P}(\pi^b))$. $\widehat{\beta} = \widehat{\Psi}_{AP,BC}(M, \mathcal{P}(\pi^b))$ can be implemented as

$$\widehat{\beta}_i = \sum_{S \subseteq \{1,\dots,n\} \setminus \{i\}} w \cdot \frac{\tilde{c}(M, \mathcal{P}(\pi^b), i)}{|S| + 1} \cdot \tilde{\Delta}_{S \cup \{i\}},$$

where $\tilde{c}(M, \mathcal{P}(\pi^b), i) = \mathbb{1}\left[\widehat{\beta}_{SV,i} > 0\right]$ with $\widehat{\beta}_{SV} = \widehat{\Psi}_{SV,BC}(M, \mathcal{P}(\pi^b))$ (see Section 4.1 for how to calculate $\widehat{\Psi}_{SV,BC}$), $w = \frac{1}{2^n - 1}$ and $\tilde{\Delta}_{S \cup \{i\}} = \min_{\pi \in \mathcal{P}'(\pi^b)} J(\pi_{S \cup \{i\}}^{*|\pi}, \pi_{-S \cup \{i\}}) - \max_{\pi \in \mathcal{P}'(\pi^b)} J(\pi)$. Here, we used the fact that c (in this case, estimate $\tilde{c}$) can be obtained via Shapley value (in this case, Blackstone consistent Shapley value).

E Experimental Setup and Implementation Details

In this section, we provide additional information on experimental setup and implementation details.

E.1 Additional Information on Experimental Setup

Environment 1: The exact penalties and rewards of the Gridworld environment (Fig. 1) are as follows: -0.01 for blank cells and S cells, -0.02 for F cells, -0.5 for H cells and $+1$ for cell G. Moreover, the cost of intervention C is -0.05. The size of the environment's state space is 64

(the state space represents cells of the Gridworld). The action space of agent A_2 is $\{0, 1\}$, which corresponds to *don't intervene* and *intervene*, and the action space of agent A_1 is $\{0, 1, 2, 3\}$, i.e. *move left*, *move right*, *move up* and *move down*. Note also that the actor remains at the same cell if it takes an action which would take it out of the environment. Finally, the specification of the personal policy of agent A_1 can be found in the source code, and more precisely in function `instantiate_behavior_policy_1()` of `env_gridworld.py`.

Environment 2: The state space of the Graph environment (Fig. 2) is defined by possible distributions of the 4 agents over the nodes of the graph, 66 states in total. The action space of each agent is $\{0, 1\}$, and the time-horizon of the environment is 5. We test multiple variants of this environment, each of which defines a different reward function. In all the variants, the reward at each time-step $t < 4$ is $+1$ if some formation constraint is satisfied and -1 if not, at time-step $t = 4$ the reward is always 0. Next, we describe in more detail the formation constraints and behavior policies for the Graph environment in the first (Coordination) and the second (Robustness) set of experiments.

Coordination: In the first set of experiments, we assign weights $w_1 = 1$, $w_2 = 2$, $w_3 = 3$ and $w_4 = 4$ to the four agents. We also consider 4 different formation constraints which are satisfied if $\sum_{i\in\{1,2,3,4\}} w_i \cdot a_i \geq h_m$, where a_i is the action taken by agent i and h_m is a threshold specific to the constraint $m \in \{1, 2, 3, 4\}$. We consider four thresholds: $h_1 = 1$, $h_2 = 7$, $h_3 = 9$ and $h_4 = 10$. For each constraint m to be satisfied, at least m number of agents need to select action 1. Each behavior ${\pi^b}_i$ takes action 0 in every state.

Robustness: In the second set of experiments, we consider one formation constraint that is satisfied if agents are arranged equally between the two levels of the graph, $\sum_{i\in\{1,2,3,4\}} a_i = 2$. When the agents are in nodes -1, 6, 7 or 8, each behavior policy ${\pi^b}_i$ takes each action with 0.5 probability. In states where agents are balanced between the levels, each behavior policy ${\pi^b}_i$ takes the action from the previous time-step with probability p_i; in unbalanced states, the action that leads to the level with the least number of agents is taken with probability p_i. We consider $p_i = 1 - (i - 1) \cdot 0.2$ for each agent $i \in \{1, 2, 3, 4\}$.

Discount factor γ is set to 0.99 in both environments.

E.2 Implementation Details

The solutions to the evaluation and optimization problems utilized by the blame attribution methods can be computed efficiently using standard (robust) optimization techniques. In the source code, the solvers of these problems are implemented as the following functions:

- policy performance evaluation in function `recursion_1_a()` (see `recursion_graph.py` and `recursion_gridworld.py`).
- problem $\arg\max_{\pi_S} J(\pi_S, {\pi^b}_{-S})$ in functions `recursion_1_c()` (see `recursion_graph.py`) and `recursion_1_c_ag1()`, `recursion_1_c_ag2()` (see `recursion_gridworld.py`).
- problem $\arg\max_{\pi\in\mathcal{P}'(\pi^b)} J(\pi)$ in function `recursion_2_a()` (see `recursion_graph.py` and `recursion_gridworld.py`).
- problem $\min_{\pi\in\mathcal{P}'(\pi^b)} J(\pi^{*|\pi}_{S\cup\{i\}}, \pi_{-S\cup\{i\}})$ in functions `recursion_3_a()` (see `recursion_graph.py`) and `recursion_3_a_ag1()`, `recursion_3_a_ag2()` (see `recursion_gridworld.py`).
- problem $\max_{\pi\in\mathcal{P}'(\pi^b)} J(\pi^{*|\pi}_{S}, \pi_{-S})$ in functions `recursion_3_b()` (see `recursion_graph.py`) and `recursion_3_b_ag1()`, `recursion_3_b_ag2()` (see `recursion_gridworld.py`).

E.3 Solutions to the Optimization Problem (P1)

(P1) might have multiple optimal solutions. Therefore, when calculating Ψ_{MER} (Section 3.1) or $\widehat{\Psi}_{MER,BC}$ (Section 4.1 and Appendix D), a way to decide which solution is going to be the blame assignment output is needed. For the experiments on the Gridworld environment the optimal solution assigning the maximum blame to A_2 was always selected. For the experiments on the Graph environment, an LP solver was applied: in the case of the Graph environment, our experiments only require the total blame assigned to the agents so any optimal solution to the LP produces the same results (see below).

L_1 Distance: For the Max-Efficient Rationality method in Fig. 3c and 3g of Section 5, we consider the L_1 distance between an output $\widehat{\beta}$ of the consistent method $\widehat{\Psi}_{MER,BC}$ and an output β of Ψ_{MER}, such that $\widehat{\beta}_i \leq \beta_i$ for all i. Notice that the L_1 distance between any two such blame assignments is equal to their difference in total blame, $\sum_{i\in\{1,\ldots,n\}} |\beta_i - \widehat{\beta}_i| = \sum_{i\in\{1,\ldots,n\}} \beta_i - \sum_{i\in\{1,\ldots,n\}} \widehat{\beta}_i$. Notice also that the total blame $\sum_{i\in\{1,\ldots,n\}} \beta_i$ (resp. $\sum_{i\in\{1,\ldots,n\}} \widehat{\beta}_i$) is the same for all optimal solutions β (resp. $\widehat{\beta}$) of (P1) with Δ_S (resp. $\tilde{\Delta}_S$), since they maximize the same objective. Hence, for obtaining the L_1 distance between the output of the consistent method $\widehat{\Psi}_{MER,BC}$ and its "targeted assignment", it suffices to compute the difference $\sum_{i\in\{1,\ldots,n\}} \beta_i - \sum_{i\in\{1,\ldots,n\}} \widehat{\beta}_i$ for any two optimal solutions β and $\widehat{\beta}$.

Total Blame: The total blame assigned by the Max-Efficient Rationality method in each of the figures 3e, 3d and 3h of Section 5 remains the same for all the optimal solutions of (P1).

E.4 Total Amount of Compute and Type of Resources

All experiments were run on a personal laptop (with Intel Core i7-8750H CPU). Experiments were also run multiple times for 10 different seeds, and we report averages and standard deviations. The total running time of the experiments on the Gridworld environment is a few minutes ($\sim$10) and of the experiments on the Graph environment a few hours ($\sim$3). Tables 2 and 3 show how much (CPU) time it takes to compute Shapley value under uncertainty (using the approaches from Section 4), for $\epsilon_{max} = \{0.01, 0.05, 0.1, 0.15, 0.2\}$. Note that Ψ_{SV} does not depend on ϵ_{max}—its running time for the Gridworld environment is 0.453125 ± 0.19111 sec and for the Graph environment is 2.02187 ± 0.06853 sec.

	$\widehat{\Psi}_{SV}$	$\widehat{\Psi}_{SV,V}$	$\widehat{\Psi}_{SV,BC}$
$\epsilon_{max} = 0.05$	0.45625 ± 0.19848	1.19843 ± 0.51044	1.38906 ± 0.60939
$\epsilon_{max} = 0.10$	0.46093 ± 0.20049	1.21093 ± 0.55609	1.45781 ± 0.71592
$\epsilon_{max} = 0.15$	0.47187 ± 0.20925	1.14062 ± 0.51864	1.45468 ± 0.66985
$\epsilon_{max} = 0.20$	0.46093 ± 0.22011	1.20937 ± 0.60934	1.72500 ± 0.96822

Table 2: Running times of different approaches for SV under uncertainty on the Gridworld environment. All times are measured in seconds (sec).

	$\widehat{\Psi}_{SV}$	$\widehat{\Psi}_{SV,V}$	$\widehat{\Psi}_{SV,BC}$
$\epsilon_{max} = 0.01$	2.06250 ± 0.10892	3.80625 ± 0.13243	92.38750 ± 1.59190
$\epsilon_{max} = 0.05$	2.07031 ± 0.18077	3.91718 ± 0.18944	91.60000 ± 1.43320
$\epsilon_{max} = 0.10$	1.97500 ± 0.05466	3.84218 ± 0.12798	92.84375 ± 3.45815

Table 3: Running times of different approaches for SV under uncertainty on the Graph environment. All times are measured in seconds (sec).

$\widehat{\Psi}_{SV,BC}$ has the largest computing time, while Ψ_{SV} and $\widehat{\Psi}_{SV}$ have the lowest computing times. These results are not surprising given that Ψ_{SV} and $\widehat{\Psi}_{SV}$ only need to compute the values once and they are not running robust optimization. Moreover, $\widehat{\Psi}_{SV,BC}$ solves $\min_{\pi\in\mathcal{P}'(\pi^b)} J(\pi^{*|\pi}_{S\cup\{i\}}, \pi_{-S\cup\{i\}})$ and $\max_{\pi\in\mathcal{P}'(\pi^b)} J(\pi^{*|\pi}_{S}, \pi_{-S})$ for each S separately, unlike $\widehat{\Psi}_{SV,V}$, which only requires robust

optimization for finding a solution to the optimization problem $\arg\max_{\pi \in \mathcal{P}'(\pi^b)} J(\pi)$. The running times of methods that compute $\widehat{\Psi}_{SV}$ do not appear to have strong dependency on ϵ_{max}. This is expected for $\widehat{\Psi}_{SV}$ since it is based on point estimates, and does not use robust optimization.

Note that the computation results obtained when calculating the aforementioned Shapley value blame assignments can be reused in computing the blame assignments of the other blame attribution methods, which we do in our experiments.

F Extended Discussion

This section of the appendix discusses different perspectives on blame attribution, and the potential negative side-effects of under-blaming agents.

F.1 Different Perspectives on Blame Attribution

Consequentialism: In this paper we follow a *consequentialist* [55] approach to the blame attribution problem, in the sense that we consider the amount of an agent's blame to depend solely on the outcome of its policy. More specifically, we consider blame attribution methods and desirable properties that measure how good or bad an agent's policy is based only on the inefficiency it causes to the multi-agent system.[9] A common objection to this type of approaches is that they do not blame an agent for violating common ground rules, i.e. they concentrate only on the ends rather than the means [56]. For example, consider an intersection accident scenario that involves two drivers: the first driver, D_1, proceeds north and the second driver D_2 proceeds east, both of them drive below the speed limit. Assume that D_2 violates a stop sign but could not do anything different to avoid the accident, while if D_1 would drive above the speed limit then with high probability the accident would have been avoided. According to consequentialism, in this example driver D_1 deserves more blame than D_2, although D_2 is the one that breaks the law.

Deontology: Consequentialism is often contrasted to another major approach in normative ethics, *deontology* [55, 57]. From a deontological perspective, the quality of an agent's policy is based on how well it follows a clear set of rules or duties[10], rather than its consequences. Therefore, a deontological approach to blame attribution would assign more blame to the second driver, from the example above, because they violate a well-known traffic regulation. Of course, deontological approaches face criticism too, for instance people argue that deontological ethics are rigid—they focus on rules, ignoring the (potentially) severe consequences of one's behavior [58]. For instance, avoiding a car crash may be more important than not violating the speed limit in the example above.

The problem of assigning blame is inherently multi-dimensional and can be viewed through both deontological and consequentialist lenses (among others). In this paper we take a consequentialist viewpoint because it provides clear and practical guidance, at least when estimating (counterfactual) outcomes is plausible. However, we do not see the two normative ethical theories as mutually exclusive [59], and thus our intention is not to replace deontological approaches, but to complement them.

F.2 Under-Blaming Agents

Apart from serving justice, blame attribution is also important for incentivizing decision makers to adopt policies that will minimize the system's inefficiency. To that end, we introduce in Section 2.3 the performance monotonicity property, the purpose of which is to motivate agents to individually improve their policies. The second property we introduce, Blackstone consistency, aims to ensure that no agent will be over-blamed when the behavior policies are not fully known to the blame attribution procedure. As expected, experimental results from Section 5 show that Blackstone consistent methods end up under-blaming agents instead. Just like over-blaming, under-blaming has its own adverse effects. Such an effect is incentivizing bad behaviors, since the agents receive reduced penalties. Therefore, there seems to be a trade-off between ensuring that no one is unjustly blamed under uncertainty and providing incentives for good behavior.

[9] This is well-aligned with the main idea of *utilitarianism* [55], which measures how good or bad an action is based only on the overall utility of its consequences.

[10] Deontology takes root from the Greek word *deon*, which means duty.

G Proofs of the Propositions from Section 3

This section of the appendix contains the proofs of the propositions from Section 3, in particular: Proposition 1, Proposition 2, and Proposition 3.

G.1 Proof of Proposition 1

Proposition 1. *Every solution to the optimization problem* (P1)*, i.e.,* Ψ_{MER}*, satisfies* $\mathcal{R}_V$ *(validity),* $\mathcal{R}_R$ *(rationality) and* $\mathcal{R}_I$ *(invariance).*

Proof. We prove the properties as follows:

- $\mathcal{R}_V$ (validity): Consider M, π^b. Every solution to the optimization problem (P1), i.e., $\beta = \Psi_{MER}(M, \pi^b)$, satisfies the constraint $\sum_{i\in\{1,...,n\}} \beta_i \leq \Delta_{\{1,...,n\}}$. The last inequality can be rewritten as $\sum_{i=1}^{n} \beta_i \leq \Delta$, and hence property $\mathcal{R}_V$ (validity) is satisfied.
- $\mathcal{R}_R$ (rationality): Consider M, π^b and $S \subseteq \{1, ..., n\}$. Every solution to the optimization problem (P1), i.e., $\beta = \Psi_{MER}(M, \pi^b)$, satisfies the constraint $\sum_{i\in S} \beta_i \leq \Delta_S$, and hence property $\mathcal{R}_R$ (rationality) is satisfied.
- $\mathcal{R}_I$ (invariance): Consider M, π^b, and an agent i such that $\Delta_{S\cup\{i\}} = \Delta_S$ for all S. This implies that $\Delta_i = \Delta_\emptyset = 0$. Now, due to the constraints of the optimization problem (P1), every solution to the optimization problem (P1), i.e., $\beta = \Psi_{MER}(M, \pi^b)$, satisfies the constraint $\beta_i \leq \Delta_i = 0$. Note also that $\sum_{j\in S} \beta_j \leq \Delta_{S\cup\{i\}} - \beta_i = \Delta_S - \beta_i$, but also $\sum_{j\in S} \beta_j \leq \Delta_S$ (where $i \notin S$). Therefore, the constraints in which agent i participates can be replaced by the the constraint $\beta_i \leq 0$. Together with the fact that the objective function is the total blame, this implies that the optimal β_i is independent of β_j ($j \neq i$), and furthermore that its value is equal to $\beta_i = 0$. Hence, property $\mathcal{R}_I$ (invariance) is satisfied.

□

G.2 Proof of Proposition 2

Proposition 2. $\Psi_{MC}(M, \pi^b) = (\Delta_1, ..., \Delta_n)$ *satisfies* $\mathcal{R}_S$ *(symmetry),* $\mathcal{R}_I$ *(invariance),* $\mathcal{R}_{CM}$ *(contribution monotonicity) and* $\mathcal{R}_{PerM}$ *(performance monotonicity).*

Proof. We prove the properties as follows:

- $\mathcal{R}_S$ (symmetry): Consider M, π^b, and agents i and j such that $\Delta_{S\cup\{i\}} = \Delta_{S\cup\{j\}}$ for all $S \subseteq \{1, ..., n\}\backslash\{i, j\}$. Notice that $\Delta_i = \Delta_j$. By using the definition of $\beta = \Psi_{MC}(M, \pi^b)$, we have that $\beta_i = \Delta_i = \Delta_j = \beta_j$. Hence, property $\mathcal{R}_S$ (symmetry) is satisfied.
- $\mathcal{R}_I$ (invariance): Consider M, π^b, and agent i such that $\Delta_{S\cup\{i\}} = \Delta_S$ for all S. Given the definition of $\beta = \Psi_{MC}(M, \pi^b)$, this implies that $\beta_i = \Delta_i = \Delta_\emptyset = 0$. Hence, property $\mathcal{R}_I$ (invariance) is satisfied.
- $\mathcal{R}_{CM}$ (contribution monotonicity): Consider M^1, ${\pi^b}^1$, M^2, ${\pi^b}^2$, and agent i such that $\Delta^1_{S\cup\{i\}} - \Delta^1_S \geq \Delta^2_{S\cup\{i\}} - \Delta^2_S$ for all S. By using the definitions of $\beta^1 = \Psi_{MC}(M^1, {\pi^b}^1)$ and $\beta^2 = \Psi_{MC}(M^2, {\pi^b}^2)$, we have that ${\beta_i}^1 = \Delta^1_i = \Delta^1_{\emptyset\cup\{i\}} - \Delta^1_\emptyset \geq \Delta^2_{\emptyset\cup\{i\}} - \Delta^2_\emptyset = \Delta^2_i = {\beta_i}^2$. Hence, property $\mathcal{R}_{CM}$ (contribution monotonicity) is satisfied.
- $\mathcal{R}_{PerM}$ (performance monotonicity): Consider M, ${\pi^b}_{-i}$, π_i and π'_i such that $J(\pi_i, {\pi^b}_{-i}) \leq J(\pi'_i, {\pi^b}_{-i})$. This implies that:

$$\begin{aligned} & J(\pi_i, {\pi^b}_{-i}) \leq J(\pi'_i, {\pi^b}_{-i}) \Rightarrow \\ \Rightarrow & J(\pi_i^{*|\pi^b}, {\pi^b}_{-i}) - J(\pi_i, {\pi^b}_{-i}) \geq J(\pi_i^{*|\pi^b}, {\pi^b}_{-i}) - J(\pi'_i, {\pi^b}_{-i}) \Rightarrow \\ \Rightarrow & \Delta_i \geq \Delta'_i. \end{aligned}$$

  By using the definitions of $\beta = \Psi_{MC}(M, (\pi_i, {\pi^b}_{-i}))$ and $\beta' = \Psi_{MC}(M, (\pi'_i, {\pi^b}_{-i}))$, we obtain that $\beta_i = \Delta_i \geq \Delta'_i = {\beta_i}'$. Hence, property $\mathcal{R}_{PerM}$ (performance monotonicity) is satisfied.

□

G.3 Proof of Proposition 3

Proposition 3. *No blame attribution method Ψ satisfies $\mathcal{R}_E$ (efficiency), $\mathcal{R}_S$ (symmetry), $\mathcal{R}_I$ (invariance) and $\mathcal{R}_{PerM}$ (performance monotonicity).*

Proof. We prove the stated impossibility result by contradiction. Suppose that there is a blame attribution method Ψ that satisfies $\mathcal{R}_E$ (efficiency), $\mathcal{R}_S$ (symmetry), $\mathcal{R}_I$ (invariance) and $\mathcal{R}_{PerM}$ (performance monotonicity).

Consider an MMDP M with two agents $\{1, 2\}$, two states—the initial state and the terminal state—and the action space $\mathcal{A} = \{0, 1, 2\} \times \{0, 1, 2\}$. In the initial state, the agents obtain zero reward when they both take action 0, reward equal to 2 when one of them takes action 0 and the other one action 2 or they both take action 2, and reward equal to 0.9 when they take any other pair of actions. After the agents perform their actions in the initial state, the MMDP transitions to the terminal state. Consider also the deterministic policies: ${\pi^b}_2$ that takes action 0, π_1 that takes action 0 and π'_1 that takes action 1, in the initial state.

We have the following three observations:

- Note that $J(\pi_1, {\pi^b}_2) \leq J(\pi'_1, {\pi^b}_2)$ and hence from property $\mathcal{R}_{PerM}$ (performance monotonicity) we have that $\beta_1 \geq \beta'_1$, where $\beta = \Psi(M, (\pi_1, {\pi^b}_2))$ and $\beta' = \Psi(M, (\pi'_1, {\pi^b}_2))$.
- Note that $\Delta_{\{1\}} = \Delta_{\{2\}} = 2$ and thus from property $\mathcal{R}_S$ (symmetry) it follows that $\beta_1 = \beta_2$. Also, from property $\mathcal{R}_E$ (efficiency) we have that $\beta_1 + \beta_2 = \Delta = 2$, and hence $\beta_1 = 1$ and $\beta_2 = 1$.
- Note that $\Delta'_{\{2\}} = 0$ and $\Delta'_{\{1,2\}} = \Delta'_{\{1\}} = 1.1$ and thus from property $\mathcal{R}_I$ (invariance) it follows that $\beta'_2 = 0$. From property $\mathcal{R}_E$ (efficiency) we have that $\beta'_1 + \beta'_2 = \Delta' = 1.1$ and hence $\beta'_1 = 1.1$, which contradicts the first two observations.

□

H Proof of Theorem 1

In this section, we provide a proof of Theorem 1. Since this proof utilizes the results of [60], we first provide some background details on these results.

H.1 Background

To prove the uniqueness result for the Shapley Value method, Theorem 1, we use a result from [60]. Before we embark on the proof, we set the necessary background. Let N be a set, such that $N \neq \emptyset$, $|N| < \infty$, and $u : 2^N \to \mathbb{R}$ be a function such that $u(\emptyset) = 0$. Then we call N set of agents and u game, and denote with $\mathcal{G}^N$ the class of games with player set N. We say that a game $u \in \mathcal{G}^N$ is *monotone*, if for each $S, T \subseteq N$, $S \subseteq T$; $u(S) \leq u(T)$. Moreover, we say that function $\psi : G \to \mathbb{R}^N$ is a solution on the class $G \in \mathcal{G}^N$. Next, we state three axioms from [60]:

- *Pareto Optimality (PO)*: We say that a solution ψ on class of games $G \subseteq \mathcal{G}^N$ satisfies *PO* (Pareto optimality), if for each game $u \in G$: $\sum_{i \in N} \psi_i(u) = u(N)$.
- *Equal Treatment Property (ETP)*: We say that a solution ψ on class of games $G \subseteq \mathcal{G}^N$ satisfies *ETP* (equal treatment property), if for each game $u \in G$ and $i, j \in N$; $\psi_i(u) = \psi_j(u)$, whenever $u(S \cup \{i\}) - u(S) = u(S \cup \{j\}) - u(S)$ for every $S \subseteq N \backslash \{i, j\}$.
- *Marginality (M)*: We say that a solution ψ on class of games $G \subseteq \mathcal{G}^N$ satisfies *M* (marginality), if for all games $u, v \in G$ and $i \in N$: $\psi_i(u) = \psi_i(v)$, whenever $u(S \cup \{i\}) - u(S) = v(S \cup \{i\}) - v(S)$ for every $S \subseteq N$.

We also define the Shapley value method for this setting. For any game $u \in \mathcal{G}^N$, the Shapley value solution ϕ is given by

$$\phi_i(u) = \sum_{S \subseteq N \backslash \{i\}} w_S \cdot [u(S \cup \{i\}) - u(S)], \tag{5}$$

where coefficients w_S are set to $w_S = \frac{|S|!(|N|-|S|-1)!}{|N|!}$.

Next we restate Theorem 3.9 from [60]:

Theorem 4. *Solution ψ defined on the class of monotone games satisfies axiom PO (Pareto optimality), ETP (equal treatment Property) and M (marginality), iff it is the Shapley value solution.*

We introduce a slightly different axiom than *M* (marginality):

- *Unequal Marginality (UM)*: We say that a solution ψ on class of games $G \subseteq \mathcal{G}^N$ satisfies *UM* (unequal marginality), if for all games $u, v \in G$ and $i \in N$: $\psi_i(u) \geq \psi_i(v)$, whenever $u(S \cup \{i\}) - u(S) \geq v(S \cup \{i\}) - v(S)$ for every $S \subseteq N$.

We also state a Corollary of Theorem 4:

Corollary 1. *Solution ψ defined on the class of monotone games satisfies axiom PO (Pareto optimality), ETP (equal treatment Property) and UM (unequal marginality), iff it is the Shapley value solution.*

Proof. We prove that Shapley value solution ϕ satisfies axiom *UM* (unequal marginality). Consider monotone games u, v, and agent $i \in N$ such that $u(S \cup \{i\}) - u(S) \geq v(S \cup \{i\}) - v(S)$ for every $S \subseteq N$, then;

$$\begin{aligned}\phi_i(u) = &\sum_{S \subseteq N \setminus \{i\}} w_S \cdot [u(S \cup \{i\}) - u(S)] \geq \\ \geq &\sum_{S \subseteq N \setminus \{i\}} w_S \cdot [v(S \cup \{i\}) - v(S)] = \\ = &\phi_i(v).\end{aligned}$$

Since *UM* (unequal marginality) is a stronger axiom than *M* (marginality), and Shapley value solution satisfies it, the uniqueness result stated in the Corollary holds because of Theorem 4. □

Consider M, π^b and notice that $\Delta_\emptyset = J(\pi^b) - J(\pi^b) = 0$. We say that set of agents N and game u are defined by M, π^b, if $N = \{1, ..., n\}$ and $u(S) = \Delta_S$ for every S. We denote with $\mathcal{H}$ the class of games that can be defined in that way. Let ψ_{SV} be the solution on class $\mathcal{H}$ such that for every M, π^b, $\psi_{SV}(u) = \Psi_{SV}(M, \pi^b)$, where game u is defined by M, π^b. Given Eq. (5), this implies that ψ_{SV} is the Shapley Value solution on $\mathcal{H}$.

We state three simple lemmas that show a one to one correspondence between the axioms *PO* (Pareto optimality), *ETP* (equal treatment property) and *UM* (unequal marginality) and blame attribution properties:

Lemma 1. *Let Ψ be a blame attribution method and ψ a solution on $\mathcal{H}$, such that for every M, π^b, $\Psi(M, \pi^b) = \psi(u)$, where game u is defined by M, π^b. Then, Ψ satisfies $\mathcal{R}_E$ (efficiency) iff ψ satisfies PO (Pareto optimality) on $\mathcal{H}$.*

Proof. Consider M, π^b and game u defined by M, π^b. Then the statement is true because $u(N) = \Delta_{\{1,...,n\}} = \Delta$. □

Lemma 2. *Let Ψ be a blame attribution method and ψ a solution on $\mathcal{H}$, such that for every M, π^b, $\Psi(M, \pi^b) = \psi(u)$, where game u is defined by M, π^b. Then, Ψ satisfies $\mathcal{R}_S$ (symmetry) iff ψ satisfies ETP (equal treatment property) on $\mathcal{H}$.*

Proof. Consider M, π^b and game u defined by M, π^b. Given that $u(S) = \Delta_S$ for every S, we have that for every i and j, $\Delta_{S \cup \{i\}} - \Delta_S = \Delta_{S \cup \{j\}} - \Delta_S$ iff $u(S \cup \{i\}) - u(S) = u(S \cup \{j\}) - u(S)$. Hence, the statement is true. □

Lemma 3. *Let Ψ be a blame attribution method and ψ a solution on $\mathcal{H}$, such that for every M, π^b, $\Psi(M, \pi^b) = \psi(u)$, where game u is defined by M, π^b. Then, Ψ satisfies $\mathcal{R}_{CM}$ (contribution monotonicity) iff ψ satisfies UM (unequal marginality) on $\mathcal{H}$.*

Proof. Consider M^1, π^{b1} and M^2, π^{b2}, and games u^1 and u^2 defined by M^1, π^{b1} and M^2, π^{b2}, respectively. Given that $u^1(S) = \Delta^1_S$ and $u^2(S) = \Delta^2_S$ for every S, we have that for every i, $\Delta^1_{S \cup \{i\}} - \Delta^1_S \geq \Delta^2_{S \cup \{i\}} - \Delta^2_S$ iff $u^1(S \cup \{i\}) - u^1(S) \geq u^2(S \cup \{i\}) - u^2(S)$. Hence, the statement is true. □

H.2 Proof

Theorem 1. *$\Psi_{SV}(M, \pi^b) = (\beta_1, ..., \beta_n)$, where β_i is defined by Eq. (1) and $w_S = \frac{|S|!(n-|S|-1)!}{n!}$, is a unique blame attribution method satisfying $\mathcal{R}_E$ (efficiency), $\mathcal{R}_S$ (symmetry) and $\mathcal{R}_{CM}$ (contribution monotonicity). Additionally, Ψ_{SV} satisfies $\mathcal{R}_V$ (validity) and $\mathcal{R}_I$ (invariance).*

Proof. Consider M, π^b and game u defined by M, π^b. Consider also S and T such that $S \subseteq T$. We have that:

$$\begin{aligned} &J(\pi_T^{*|\pi^b}, {\pi^b}_{-T}) \geq J(\pi_S^{*|\pi^b}, {\pi^b}_{-S}) \Rightarrow \\ \Rightarrow &J(\pi_T^{*|\pi^b}, {\pi^b}_{-T}) - J(\pi^b) \geq J(\pi_S^{*|\pi^b}, {\pi^b}_{-S}) - J(\pi^b) \Rightarrow \\ \Rightarrow &\Delta_T \geq \Delta_S \Rightarrow u(T) \geq u(S). \end{aligned}$$

This implies that class $\mathcal{H}$ consists only of monotone games, and hence by Corollary 1 we have that ψ_{SV} is a unique solution on $\mathcal{H}$ satisfying *PO* (Pareto optimality), *ETP* (equal treatment property) and *UM* (unequal marginality). Given Lemmas 1, 2, and 3, this implies that Ψ_{SV} is a unique blame attribution method satisfying $\mathcal{R}_E$ (efficiency), $\mathcal{R}_S$ (symmetry) and $\mathcal{R}_{CM}$ (contribution monotonicity).

We also prove the properties $\mathcal{R}_V$ (validity) and $\mathcal{R}_I$ (invariance) as follows:

- $\mathcal{R}_V$ (validity): Consider M, π^b. Given that Ψ_{SV} satisfies property $\mathcal{R}_E$ (efficiency), it holds that $\sum_{i \in \{1,...,n\}} \beta_i = \Delta$. Hence, property $\mathcal{R}_V$ (validity) is satisfied.
- $\mathcal{R}_I$ (invariance): Consider M, π^b, and agent i such that $\Delta_{S \cup \{i\}} = \Delta_S$ for all S. This implies that $J(\pi_{S \cup \{i\}}^{*|\pi^b}, {\pi^b}_{-S \cup \{i\}}) = J(\pi_S^{*|\pi^b}, {\pi^b}_{-S})$ for all S. Given the definition of $\Psi_{SV}(M, \pi^b)$, we have that $\beta_i = 0$. Hence, property $\mathcal{R}_I$ (invariance) is satisfied.

□

I Proof of Theorem 2

Before we proceed with the proof of Theorem 2, notice that the contribution function c from Section 3.4 can be rewritten in the equivalent form:

$$c(M, \pi^b, i) = \begin{cases} 0 & \text{if } \Delta_{S \cup \{i\}} = \Delta_S, \quad \forall S \subseteq \{1, ..., n\} \\ 1 & \text{otherwise} \end{cases}.$$

We also state the following lemmas:

Lemma 4. *Consider a function $f : 2^{\{1,...,n\}} \to \mathbb{R}_{\geq 0}$. There exist some MMDP M and agents' behavior joint policy π^b such that the marginal inefficiency of every subset of agents S is equal to $f(S)$, iff $f(\emptyset) = 0$ and $f(S_1) \leq f(S_2)$ whenever $S_1 \subseteq S_2$, where S_1 and S_2 are subsets of $\{1, ..., n\}$.*

Proof. First, we show that the conditions on function f are necessary:

- Suppose that there exist M, π^b such that $\Delta_\emptyset > 0$. Given the definition of marginal inefficiency this would imply that $J(\pi^b) > J(\pi^b)$. Hence, we reach a contradiction.
- Suppose that there exist M, π^b such that $\Delta_{S_1} > \Delta_{S_2}$, where $S1 \subseteq S_2$. Given the definition of marginal inefficiency this would imply that $J(\pi_{S_1}^{*|\pi^b}, {\pi^b}_{-S_1}) > J(\pi_{S_2}^{*|\pi^b}, {\pi^b}_{-S_2})$. Hence, we reach a contradiction.

Next we show that the conditions on function f are sufficient. Consider an MMDP M with two states—the initial state and the terminal state—and the action space $\mathcal{A} = \times_{i=1}^n \{0, 1\}$. In the initial state, the agents obtain zero reward when they all take action 0 and reward $f(S)$ when agents in S take action 1 and the rest of the agents take action 0. Consider also the deterministic joint policy π^b, where every agent takes action 0. Notice that $J(\pi^b) = 0$.

For every subset of agents S it holds that $J(\pi_S^{*|\pi^b}, {\pi^b}_{-S}) = f(S)$, because taking action 1 is the best that every agent in S can do. Hence, for the marginal inefficiency of S we have that $\Delta_S = J(\pi_S^{*|\pi^b}, {\pi^b}_{-S}) - J(\pi^b) = f(S)$. □

Lemma 5. *Let Ψ satisfy $\mathcal{R}_{cParM}$ (c-participation monotonicity). Then, for every M^1, ${\pi^b}^1$ and M^2, ${\pi^b}^2$ such that $c(M^1, {\pi^b}^1, i) = c(M^2, {\pi^b}^2, i)$ for every i, ${\beta_i}^1 = {\beta_i}^2$ whenever $\Delta^1_{S\cup\{i\}} = \Delta^2_{S\cup\{i\}}$ for all S, where $\beta^1 = \Psi(M^1, {\pi^b}^1)$ and $\beta^2 = \Psi(M^2, {\pi^b}^2)$.*

Proof. Consider agent i such that $\Delta^1_{S\cup\{i\}} = \Delta^2_{S\cup\{i\}}$ for all S. Given that Ψ satisfies $\mathcal{R}_{cParM}$ (c-participation monotonicity), this implies ${\beta_i}^1 \geq {\beta_i}^2$ and ${\beta_i}^1 \leq {\beta_i}^2$, and hence ${\beta_i}^1 = {\beta_i}^2$. □

Lemma 6. *Let Ψ satisfy $\mathcal{R}_{RcParM}$ (relative c-participation monotonicity). Then, for every M^1, ${\pi^b}^1$ and M^2, ${\pi^b}^2$ such that $c(M^1, {\pi^b}^1, i) = c(M^2, {\pi^b}^2, i)$ for every i, ${\beta_j}^1 - {\beta_j}^2 = {\beta_k}^1 - {\beta_k}^2$ whenever $c(M^1, {\pi^b}^1, j) = c(M^1, {\pi^b}^1, k)$ and $\Delta^1_{S\cup\{j\}} - \Delta^2_{S\cup\{j\}} = \Delta^1_{S\cup\{k\}} - \Delta^2_{S\cup\{k\}}$ for every $S \subseteq \{1, ..., n\}\backslash\{j, k\}$, where $\beta^1 = \Psi(M^1, {\pi^b}^1)$ and $\beta^2 = \Psi(M^2, {\pi^b}^2)$.*

Proof. Consider agents j and k such that $c(M^1, {\pi^b}^1, j) = c(M^1, {\pi^b}^1, k)$ and $\Delta^1_{S\cup\{j\}} - \Delta^2_{S\cup\{j\}} = \Delta^1_{S\cup\{k\}} - \Delta^2_{S\cup\{k\}}$ for all $S \subseteq \{1, ..., n\}\backslash\{j, k\}$. Given that Ψ satisfies $\mathcal{R}_{RcParM}$ (relative c-participation monotonicity), this implies ${\beta_j}^1 - {\beta_j}^2 \geq {\beta_k}^1 - {\beta_k}^2$ and ${\beta_j}^1 - {\beta_j}^2 \leq {\beta_k}^1 - {\beta_k}^2$, and hence ${\beta_j}^1 - {\beta_j}^2 = {\beta_k}^1 - {\beta_k}^2$. □

Proof of Theorem 2

Theorem 2. *$\Psi_{AP}(M, \pi^b) = (\beta_1, ..., \beta_n)$, where β_i is defined by Eq. (2) and $w = \frac{1}{2^n - 1}$, is a unique blame attribution method that satisfies $\mathcal{R}_{AE}$ (average-efficiency), $\mathcal{R}_S$ (symmetry), $\mathcal{R}_I$ (invariance), $\mathcal{R}_{cParM}$ (c-participation monotonicity) and $\mathcal{R}_{RcParM}$ (relative c-participation monotonicity). Furthermore, Ψ_{AP} satisfies $\mathcal{R}_{cPerM}$ (c-performance monotonicity) and $\mathcal{R}_V$ (validity).*

Proof. The proof is separated into two parts. In the first part we prove that Ψ_{AP} satisfies the mentioned properties, while in the second part we show that if a blame attribution method satisfies all mentioned properties, it must be the Ψ_{AP} method.

First Part

We prove the properties as follows:

- $\mathcal{R}_{AE}$ (average-efficiency): Consider M, π^b. By using the definition of $\beta = \Psi_{AP}(M, \pi^b)$:

$$\begin{aligned}
\sum_{i=1}^{n} \beta_i &= \sum_{i=1}^{n} \sum_{S\subseteq\{1,...,n\}\backslash\{i\}} w \cdot \frac{c(M, \pi^b, i)}{\sum_{j\in S} c(M, \pi^b, j) + 1} \cdot \Delta_{S\cup\{i\}} = \\
&= \frac{1}{2^n - 1} \cdot \sum_{i\in\{1,...,n\}|c(M,\pi^b,i)=1} \sum_{S\subseteq\{1,...,n\}\backslash\{i\}} \frac{1}{\sum_{j\in S} c(M, \pi^b, j) + 1} \cdot \Delta_{S\cup\{i\}} = \\
&= \frac{1}{2^n - 1} \cdot \sum_{i\in\{1,...,n\}|c(M,\pi^b,i)=1} \sum_{S\subseteq\{1,...,n\}|i\in S} \frac{1}{\sum_{j\in S} c(M, \pi^b, j)} \cdot \Delta_S = \\
&= \frac{1}{2^n - 1} \cdot \sum_{S\subseteq\{1,...,n\}} \sum_{i\in S|c(M,\pi^b,i)=1} \frac{1}{\sum_{j\in S} c(M, \pi^b, j)} \cdot \Delta_S = \\
&= \frac{1}{2^n - 1} \cdot \sum_{S\subseteq\{1,...,n\}} \Delta_S,
\end{aligned}$$

  and hence property $\mathcal{R}_{AE}$ (average-efficiency) is satisfied.
- $\mathcal{R}_S$ (symmetry): Consider M, π^b, and agents i and j such that $\Delta_{S\cup\{i\}} = \Delta_{S\cup\{j\}}$ for all $S \subseteq \{1, ..., n\}\backslash\{i, j\}$. Notice that if $\Delta_{S\cup\{i\}} = \Delta_S$ for all S then $\Delta_{S\cup\{j\}} = \Delta_S$ and $\Delta_{S\cup\{i,j\}} = \Delta_{S\cup\{j\}} = \Delta_{S\cup\{i\}}$ for every $S \subseteq \{1, ..., n\}\backslash\{i, j\}$, and hence $\Delta_{S\cup\{j\}} = \Delta_S$ for all S. Given the definition of contribution function c, this implies that if $c(M, \pi^b, i) = 0$, then $c(M, \pi^b, j) = 0$.

For similar reasons, it also holds that if $c(M, \pi^b, j) = 0$, then $c(M, \pi^b, i) = 0$, and hence $c(M, \pi^b, i) = c(M, \pi^b, j)$. By using the definition of $\beta = \Psi_{AP}(M, \pi^b)$, we have that:

$$\begin{aligned}
\beta_i &= \sum_{S \subseteq \{1,\dots,n\} \setminus \{i\}} w \cdot \frac{c(M, \pi^b, i)}{\sum_{k \in S} c(M, \pi^b, k) + 1} \cdot \Delta_{S \cup \{i\}} = \\
&= \sum_{S \subseteq \{1,\dots,n\} \setminus \{i,j\}} w \cdot \frac{c(M, \pi^b, i)}{\sum_{k \in S} c(M, \pi^b, k) + 1} \cdot \Delta_{S \cup \{i\}} + \\
&\quad + \sum_{S \subseteq \{1,\dots,n\} \setminus \{i,j\}} w \cdot \frac{c(M, \pi^b, i)}{\sum_{k \in S \cup \{j\}} c(M, \pi^b, k) + 1} \cdot \Delta_{S \cup \{i,j\}} = \\
&= \sum_{S \subseteq \{1,\dots,n\} \setminus \{i,j\}} w \cdot \frac{c(M, \pi^b, j)}{\sum_{k \in S} c(M, \pi^b, k) + 1} \cdot \Delta_{S \cup \{j\}} + \\
&\quad + \sum_{S \subseteq \{1,\dots,n\} \setminus \{i,j\}} w \cdot \frac{c(M, \pi^b, j)}{\sum_{k \in S \cup \{i\}} c(M, \pi^b, k) + 1} \cdot \Delta_{S \cup \{i,j\}} = \\
&= \sum_{S \subseteq \{1,\dots,n\} \setminus \{j\}} w \cdot \frac{c(M, \pi^b, j)}{\sum_{k \in S} c(M, \pi^b, k) + 1} \cdot \Delta_{S \cup \{j\}} = \beta_j,
\end{aligned}$$

and hence property $\mathcal{R}_S$ (symmetry) is satisfied.

- $\mathcal{R}_I$ (invariance): Consider M, π^b, and agent i such that $\Delta_{S \cup \{i\}} = \Delta_S$ for all S. Given the definitions of contribution function c and $\beta = \Psi_{AP}(M, \pi^b)$, this implies that $\beta_i = 0$. Hence, property $\mathcal{R}_I$ (invariance) is satisfied.
- $\mathcal{R}_{cParM}$ (c-participation monotonicity): Consider M^1, π^{b^1} and M^2, π^{b^2} such that $c(M^1, \pi^{b^1}, i) = c(M^2, \pi^{b^2}, i)$ for every i. Consider also agent i such that $\Delta^1_{S \cup \{i\}} \geq \Delta^2_{S \cup \{i\}}$ for all S. By using the definitions of $\beta^1 = \Psi_{AP}(M^1, \pi^{b^1})$ and $\beta^2 = \Psi_{AP}(M^2, \pi^{b^2})$, this implies:

$$\begin{aligned}
\beta^1_i &= \sum_{S \subseteq \{1,\dots,n\} \setminus \{i\}} w \cdot \frac{c(M^1, \pi^{b^1}, i)}{\sum_{j \in S} c(M^1, \pi^{b^1}, j) + 1} \cdot \Delta^1_{S \cup \{i\}} = \\
&= \sum_{S \subseteq \{1,\dots,n\} \setminus \{i\}} w \cdot \frac{c(M^2, \pi^{b^2}, i)}{\sum_{j \in S} c(M^2, \pi^{b^2}, j) + 1} \cdot \Delta^1_{S \cup \{i\}} \geq \\
&\geq \sum_{S \subseteq \{1,\dots,n\} \setminus \{i\}} w \cdot \frac{c(M^2, \pi^{b^2}, i)}{\sum_{j \in S} c(M^2, \pi^{b^2}, j) + 1} \cdot \Delta^2_{S \cup \{i\}} = \beta^2_i,
\end{aligned}$$

and hence property $\mathcal{R}_{cParM}$ (c-participation monotonicity) is satisfied.

- $\mathcal{R}_{RcParM}$ (relative c-participation monotonicity): Consider M^1, π^{b^1} and M^2, π^{b^2} such that $c(M^1, \pi^{b^1}, i) = c(M^2, \pi^{b^2}, i)$ for every i. Consider also agents j and k such that $c(M^1, \pi^{b^1}, j) = c(M^1, \pi^{b^1}, k)$ and $\Delta^1_{S \cup \{j\}} - \Delta^2_{S \cup \{j\}} \geq \Delta^1_{S \cup \{k\}} - \Delta^2_{S \cup \{k\}}$ for all $S \subseteq \{1, ..., n\} \setminus \{j, k\}$. By using the definitions of $\beta^1 = \Psi_{AP}(M^1, \pi^{b^1})$ and $\beta^2 = \Psi_{AP}(M^2, \pi^{b^2})$, this implies:

$$\begin{aligned}
{\beta_j}^1 - {\beta_j}^2 &= \sum_{S \subseteq \{1,\dots,n\} \setminus \{j\}} w \cdot \frac{c(M^1, \pi^{b^1}, j)}{\sum_{i \in S} c(M^1, \pi^{b^1}, i) + 1} \cdot \left[\Delta^1_{S \cup \{j\}} - \Delta^2_{S \cup \{j\}}\right] = \\
&= \sum_{S \subseteq \{1,\dots,n\} \setminus \{j,k\}} w \cdot \frac{c(M^1, \pi^{b^1}, j)}{\sum_{i \in S} c(M^1, \pi^{b^1}, i) + 1} \cdot \left[\Delta^1_{S \cup \{j\}} - \Delta^2_{S \cup \{j\}}\right] + \\
&\quad + \sum_{S \subseteq \{1,\dots,n\} \setminus \{j,k\}} w \cdot \frac{c(M^1, \pi^{b^1}, j)}{\sum_{i \in S \cup \{k\}} c(M^1, \pi^{b^1}, i) + 1} \cdot \left[\Delta^1_{S \cup \{j,k\}} - \Delta^2_{S \cup \{j,k\}}\right] \geq \\
&\geq \sum_{S \subseteq \{1,\dots,n\} \setminus \{j,k\}} w \cdot \frac{c(M^1, \pi^{b^1}, k)}{\sum_{i \in S} c(M^1, \pi^{b^1}, i) + 1} \cdot \left[\Delta^1_{S \cup \{k\}} - \Delta^2_{S \cup \{k\}}\right] +
\end{aligned}$$

$$+\sum_{S\subseteq\{1,\ldots,n\}\setminus\{j,k\}} w\cdot\frac{c(M^1,{\pi^b}^1,k)}{\sum_{i\in S\cup\{j\}}c(M^1,{\pi^b}^1,i)+1}\cdot\left[\Delta^1_{S\cup\{j,k\}}-\Delta^2_{S\cup\{j,k\}}\right]=$$

$$=\sum_{S\subseteq\{1,\ldots,n\}\setminus\{k\}} w\cdot\frac{c(M^1,{\pi^b}^1,k)}{\sum_{i\in S}c(M^1,{\pi^b}^1,i)+1}\cdot\left[\Delta^1_{S\cup\{k\}}-\Delta^2_{S\cup\{k\}}\right]={\beta_k}^1-{\beta_k}^2,$$

and hence property $\mathcal{R}_{RcParM}$ (relative c-participation monotonicity) is satisfied.

- $\mathcal{R}_{cPerM}$ (c-performance monotonicity): Consider M, ${\pi^b}_{-i}$, π_i and π_i' such that $J(\pi_i,{\pi^b}_{-i})\leq J(\pi_i',{\pi^b}_{-i})$ and $c(M,(\pi_i,{\pi^b}_{-i}),j)=c(M,(\pi_i',{\pi^b}_{-i}),j)$ for every j. This implies that:

$$J(\pi_i,{\pi^b}_{-i})\leq J(\pi_i',{\pi^b}_{-i})\Rightarrow$$
$$\Rightarrow J(\pi^{*|\pi^b}_{S\cup\{i\}},{\pi^b}_{-S\cup\{i\}})-J(\pi_i,{\pi^b}_{-i})\geq J(\pi^{*|\pi^b}_{S\cup\{i\}},{\pi^b}_{-S\cup\{i\}})-J(\pi_i',{\pi^b}_{-i})\Rightarrow$$
$$\Rightarrow\Delta_{S\cup\{i\}}\geq\Delta'_{S\cup\{i\}}$$

for every $S\subseteq\{1,...,n\}\setminus\{i\}$. Given the definitions of $\beta=\Psi_{AP}(M,(\pi_i,{\pi^b}_{-i}))$ and $\beta'=\Psi_{AP}(M,(\pi_i',{\pi^b}_{-i}))$, this implies that ${\beta_i}\geq{\beta_i}'$. Hence, property $\mathcal{R}_{cPerM}$ (c-performance monotonicity) is satisfied.

- $\mathcal{R}_V$ (validity): Consider M , π^b. Notice that $\sum_{S\subseteq\{1,\ldots,n\}}\frac{1}{2^n-1}\cdot\Delta_S\leq\Delta$. Given that Ψ_{AP} satisfies property $\mathcal{R}_{AE}$ (average efficiency), we have that $\sum_{i=1}^n\beta_i=\sum_{S\subseteq\{1,\ldots,n\}}\frac{1}{2^n-1}\cdot\Delta_S$, and thus $\sum_{i=1}^n\beta_i\leq\Delta$, where $\beta=\Psi_{AP}(M,\pi^b)$. Hence property $\mathcal{R}_V$ (validity) is satisfied.

Second Part

We begin by introducing some additional notation. Consider M , π^b. We define the sets of agents $C_0=\left\{i\in\{1,\ldots,n\}:c(M,\pi^b,i)=0\right\}$ and $C_1=\left\{i\in\{1,\ldots,n\}:c(M,\pi^b,i)=1\right\}$. Consider M^ϵ, ${\pi^b}^\epsilon$ such that:

$$\Delta^\epsilon_S=\begin{cases}\Delta_S+\epsilon & \text{if } S\cap C_1\neq\emptyset\\ \Delta_S & \text{otherwise,}\end{cases}\tag{6}$$

where $\epsilon>0$. Note that for every subset S such that $S\cap C_1=\emptyset$ it holds that $\Delta^\epsilon_S=0$, but we use $\Delta^\epsilon_S=\Delta_S$ for notational simplicity. Moreover, notice that Eq. (6) satisfies the conditions of Lemma 4, and hence M^ϵ, ${\pi^b}^\epsilon$ exist.

We prove that Ψ_{AP} uniquely satisfies the properties mentioned in Theorem 2 through two intermediate lemmas. Lemma 7 states that if $\Psi(M^\epsilon,{\pi^b}^\epsilon)=\Psi_{AP}(M^\epsilon,{\pi^b}^\epsilon)$ then $\Psi(M,\pi^b)=\Psi_{AP}(M,\pi^b)$, and Lemma 8 states that $\Psi(M^\epsilon,{\pi^b}^\epsilon)=\Psi_{AP}(M^\epsilon,{\pi^b}^\epsilon)$.

Lemma 7. *Consider M, π^b and M^ϵ, ${\pi^b}^\epsilon$, where Δ^ϵ_S is defined by Eq.* (6)*. If Ψ satisfies properties $\mathcal{R}_{AE}$, $\mathcal{R}_S$, $\mathcal{R}_I$, $\mathcal{R}_{cParM}$ and $\mathcal{R}_{RcParM}$ and $\Psi(M^\epsilon,{\pi^b}^\epsilon)=\Psi_{AP}(M^\epsilon,{\pi^b}^\epsilon)$, then $\Psi(M,\pi^b)=\Psi_{AP}(M,\pi^b)$.*

Proof. We state three claims that we prove after the end of the proof of Theorem 2:

Claim 1. *$c(M,\pi^b,i)=c(M^\epsilon,{\pi^b}^\epsilon,i)$ for every i.*

Claim 2. *${\beta_i}=0$ and ${\beta_i}^\epsilon=0$ for every $i\in C_0$, where $\beta=\Psi(M,\pi^b)$ and $\beta^\epsilon=\Psi(M^\epsilon,{\pi^b}^\epsilon)$.*

Claim 3. *${\beta_i}^\epsilon-\beta_i=r$ for every $i\in C_1$, where $\beta=\Psi(M,\pi^b)$ and $\beta^\epsilon=\Psi(M^\epsilon,{\pi^b}^\epsilon)$, and $r=\frac{1}{|C_1|}\cdot\sum_{S\subseteq\{1,\ldots,n\}}w\cdot[\Delta^\epsilon_S-\Delta_S]$.*

Given Claim 3, the assumption $\Psi(M^\epsilon,{\pi^b}^\epsilon)=\Psi_{AP}(M^\epsilon,{\pi^b}^\epsilon)$ implies that for every $i\in C_1$:

$$\beta_i=\sum_{S\subseteq\{1,\ldots,n\}\setminus\{i\}}w\cdot\frac{1}{\sum_{j\in S}c(M^\epsilon,{\pi^b}^\epsilon,j)+1}\cdot\Delta^\epsilon_{S\cup\{i\}}-\frac{1}{|C_1|}\cdot\sum_{S\subseteq\{1,\ldots,n\}}w\cdot[\Delta^\epsilon_S-\Delta_S]\,.\tag{7}$$

Combining Claim 2 and Eq. (7) implies that:

$$\beta_i = c(M, \pi^b, i) \cdot \left[\sum_{S \subseteq \{1,\dots,n\} \setminus \{i\}} w \cdot \frac{1}{\sum_{j \in S} c(M^\epsilon, \pi^{b^\epsilon}, j) + 1} \cdot \Delta^\epsilon_{S \cup \{i\}} - \frac{1}{|C_1|} \cdot \sum_{S \subseteq \{1,\dots,n\}} w \cdot [\Delta^\epsilon_S - \Delta_S] \right]. \tag{8}$$

Notice that $\beta = \Psi(M, \pi^b)$ is uniquely defined by the properties of Ψ, Eq. (8), and since Ψ_{AP} satisfies all properties assumed for Ψ (see Part 1), it must hold that $\Psi(M, \pi^b) = \Psi_{AP}(M, \pi^b)$. This concludes the proof of Lemma 7. □

Lemma 8. *Consider M, π^b and M^ϵ, π^{b^ϵ}, where Δ^ϵ_S is defined by Eq. (6). If Ψ satisfies properties $\mathcal{R}_{AE}$, $\mathcal{R}_S$, $\mathcal{R}_I$, $\mathcal{R}_{cParM}$ and $\mathcal{R}_{RcParM}$, then $\Psi(M^\epsilon, \pi^{b^\epsilon}) = \Psi_{AP}(M^\epsilon, \pi^{b^\epsilon})$.*

Proof. Let $I = \{1, ..., 2^{|C_1|} - 1\}$ be an index set, and for each $\iota \in I$, let S_ι be a subset of C_1 other than $\emptyset$. We assume that the indexing of subsets $S \subseteq C_1$ satisfies the following condition: for every $\iota, \zeta \in I$, $\iota < \zeta$ whenever $|S_\iota| > |S_\zeta|$.

Consider M, π^b and M^ϵ, π^{b^ϵ}, where Δ^ϵ_S is defined by Eq. (6). For each index number $\iota \in I$ consider M^ι, π^{b^ι} such that:

$$\Delta^\iota_S = \begin{cases} \epsilon & \text{if } S \cap C_1 = S_\zeta, \text{ where } \zeta > \iota \\ \Delta_S & \text{if } S \cap C_1 = \emptyset \\ \Delta_S + \epsilon & \text{otherwise,} \end{cases} \tag{9}$$

where $\epsilon > 0$. Note that for every subset S such that $S \cap C_1 = \emptyset$ it holds that $\Delta^\iota_S = 0$, but we use $\Delta^\iota_S = \Delta_S$ for notational simplicity. Moreover, notice that for every $\iota \in I$ Eq. (9) satisfies the conditions of Lemma 4, and hence M^ι, π^{b^ι} exist. Notice also that $\Delta^{2^{|C_1|}-1}_S = \Delta^\epsilon_S$ for every S.

We state four claims that we prove after the end of the proof of Theorem 2:

Claim 4. *For each $\iota \in I$, $c(M, \pi^b, i) = c(M^\iota, \pi^{b^\iota}, i)$ for every i.*

Claim 5. *For each $\iota \in I$, ${\beta_i}^\iota = 0$ for every $i \in C_0$, where $\beta^\iota = \Psi(M^\iota, \pi^{b^\iota})$.*

Claim 6. *For each $\iota \in I \setminus \{2^{|C_1|} - 1\}$, ${\beta_i}^{\iota+1} - {\beta_i}^\iota = 0$ for every $i \in C_1 \setminus S_{\iota+1}$, where $\beta^\iota = \Psi(M^\iota, \pi^{b^\iota})$ and $\beta^{\iota+1} = \Psi(M^{\iota+1}, \pi^{b^{\iota+1}})$.*

Claim 7. *For each $\iota \in I \setminus \{2^{|C_1|} - 1\}$, ${\beta_i}^{\iota+1} - {\beta_i}^\iota = r$ for every $i \in S_{\iota+1}$, where $\beta^\iota = \Psi(M^\iota, \pi^{b^\iota})$ and $\beta^{\iota+1} = \Psi(M^{\iota+1}, \pi^{b^{\iota+1}})$, and $r = \frac{1}{|S_{\iota+1}|} \cdot \sum_{S \subseteq \{1,\dots,n\}} w \cdot \left[\Delta^{\iota+1}_S - \Delta^\iota_S\right]$.*

We prove that $\Psi(M^{2^{|C_1|}-1}, \pi^{b^{2^{|C_1|}-1}}) = \Psi_{AP}(M^{2^{|C_1|}-1}, \pi^{b^{2^{|C_1|}-1}})$, by using induction in the index number ι. Note that because $\Delta^{2^{|C_1|}-1}_S = \Delta^\epsilon_S$ for every S, showing $\Psi(M^{2^{|C_1|}-1}, \pi^{b^{2^{|C_1|}-1}}) = \Psi_{AP}(M^{2^{|C_1|}-1}, \pi^{b^{2^{|C_1|}-1}})$ is equivalent to showing that $\Psi(M^\epsilon, \pi^{b^\epsilon}) = \Psi_{AP}(M^\epsilon, \pi^{b^\epsilon})$.

$\boldsymbol{\iota = 1}$: We show that $\Psi(M^1, \pi^{b^1}) = \Psi_{AP}(M^1, \pi^{b^1})$. Because of the condition that the indexing of the subsets of C_1 has to satisfy, it follows that $S_1 = C_1$. Notice that for every two agents $i, j \in C_1$ it holds that $S \cup \{i\} \cap C_1 \neq C_1 = S_1$ and $S \cup \{j\} \cap C_1 \neq C_1 = S_1$, for every $S \subseteq \{1, ..., n\} \setminus \{i, j\}$. By using Eq. (9), this implies that $\Delta^1_{S \cup \{i\}} = \epsilon = \Delta^1_{S \cup \{j\}}$, for every $S \subseteq \{1, ..., n\} \setminus \{i, j\}$. Given that Ψ is assumed to satisfy $\mathcal{R}_S$ (symmetry), this implies that ${\beta_i}^1 = {\beta_j}^1$, where $\beta^1 = \Psi(M^1, \pi^{b^1})$. It follows that for every $i \in C_1$:

$${\beta_i}^1 = \frac{1}{|C_1|} \cdot \sum_{j \in C_1} {\beta_j}^1.$$

By using Claim 5, we have that ${\beta_i}^1 = \frac{1}{|C_1|} \cdot \sum_{j \in \{1,\dots,n\}} {\beta_j}^1$. Given that Ψ satisfies $\mathcal{R}_{AE}$ (average efficiency), this implies that for every $i \in C_1$:

$${\beta_i}^1 = \frac{1}{|C_1|} \cdot \sum_{j \in \{1,\dots,n\}} {\beta_j}^1 = \frac{1}{|C_1|} \cdot \sum_{S \subseteq \{1,\dots,n\}} w \cdot \Delta^1_S. \tag{10}$$

Combining Claim 5 and Eq. (10) implies that:

$$\beta_i{}^1 = c(M, \pi^b, i) \cdot \frac{1}{|C_1|} \cdot \sum_{S \subseteq \{1,\ldots,n\}} w \cdot \Delta_S^1. \tag{11}$$

Notice that $\beta^1 = \Psi(M^1, \pi^{b^1})$ is uniquely defined by the properties of Ψ, Eq. (11), and since Ψ_{AP} satisfies all properties assumed for Ψ (see Part 1), it must hold that $\Psi_{AP}(M^1, \pi^{b^1}) = \beta^1$, and hence $\Psi(M^1, \pi^{b^1}) = \Psi_{AP}(M^1, \pi^{b^1})$.

$\boldsymbol{\iota \in I \backslash \{2^{|C_1|} - 1\}}$: Given that $\Psi(M^\iota, \pi^{b^\iota}) = \Psi_{AP}(M^\iota, \pi^{b^\iota})$, we show that $\Psi(M^{\iota+1}, \pi^{b^{\iota+1}}) = \Psi_{AP}(M^{\iota+1}, \pi^{b^{\iota+1}})$.

By using the definition of $\Psi_{AP}(M^\iota, \pi^{b^\iota})$ and Claim 6, the assumption $\Psi(M^\iota, \pi^{b^\iota}) = \Psi_{AP}(M^\iota, \pi^{b^\iota})$ implies that for every $i \in C_1 \backslash S_{\iota+1}$:

$$\beta_i{}^{\iota+1} = \beta_i{}^\iota = \sum_{S \subseteq \{1,\ldots,n\} \backslash \{i\}} w \cdot \frac{1}{\sum_{j \in S} c(M^\iota, \pi^{b^\iota}, j) + 1} \cdot \Delta_{S \cup \{i\}}^\iota. \tag{12}$$

By using the definition of $\Psi_{AP}(M^\iota, \pi^{b^\iota})$ and Claim 7, the assumption $\Psi(M^\iota, \pi^{b^\iota}) = \Psi_{AP}(M^\iota, \pi^{b^\iota})$ implies that for every $i \in S_{\iota+1}$:

$$\begin{aligned}\beta_i{}^{\iota+1} =& \beta_i{}^\iota + \frac{1}{|S_{\iota+1}|} \cdot \sum_{S \subseteq \{1,\ldots,n\}} w \cdot \left[\Delta_S^{\iota+1} - \Delta_S^\iota\right] = \\ =& \sum_{S \subseteq \{1,\ldots,n\} \backslash \{i\}} w \cdot \frac{1}{\sum_{j \in S} c(M^\iota, \pi^{b^\iota}, j) + 1} \cdot \Delta_{S \cup \{i\}}^\iota + \frac{1}{|S_{\iota+1}|} \cdot \sum_{S \subseteq \{1,\ldots,n\}} w \cdot \left[\Delta_S^{\iota+1} - \Delta_S^\iota\right].\end{aligned} \tag{13}$$

Notice that $\beta^{\iota+1} = \Psi(M^{\iota+1}, \pi^{b^{\iota+1}})$ is uniquely defined by properties of Ψ, Claim 5, Eq. (12) and Eq. (13), and since Ψ_{AP} satisfies all the properties assumed for Ψ (see Part 1), it must hold that $\Psi(M^{\iota+1}, \pi^{b^{\iota+1}}) = \beta^{\iota+1}$, and hence $\Psi(M^{\iota+1}, \pi^{b^{\iota+1}}) = \Psi_{AP}(M^{\iota+1}, \pi^{b^{\iota+1}})$. This concludes the induction step and the proof of Lemma 8. □

The second part of the proof is hence concluded. □

Proofs of the Claims 1, 2 and 3

Claim 1. *$c(M, \pi^b, i) = c(M^\epsilon, \pi^{b^\epsilon}, i)$ for every i.*

Proof. Consider agent i such that $i \in C_1$. Given Eq. (6), this implies that $\Delta_i^\epsilon = \Delta_i + \epsilon > 0 = \Delta_\emptyset^\epsilon$, and thus $c(M^\epsilon, \pi^{b^\epsilon}, i) = 1$. Hence, $c(M, \pi^b, i) = c(M^\epsilon, \pi^{b^\epsilon}, i)$.

Consider agent i such that $i \in C_0$. Given the definition of contribution function c, we have that $\Delta_{S \cup \{i\}} = \Delta_S$ for every S. By using Eq. (6), this implies that $\Delta_{S \cup \{i\}}^\epsilon = \Delta_S^\epsilon$ for every S such that $S \cap C_1 = \emptyset$ and $\Delta_{S \cup \{i\}}^\epsilon = \Delta_{S \cup \{i\}} + \epsilon = \Delta_S + \epsilon = \Delta_S^\epsilon$ for every S such that $S \cap C_1 \neq \emptyset$, and thus $c(M^\epsilon, \pi^{b^\epsilon}, i) = 0$. Hence, $c(M, \pi^b, i) = c(M^\epsilon, \pi^{b^\epsilon}, i)$. □

Claim 2. *$\beta_i = 0$ and $\beta_i{}^\epsilon = 0$ for every $i \in C_0$, where $\beta = \Psi(M, \pi^b)$ and $\beta^\epsilon = \Psi(M^\epsilon, \pi^{b^\epsilon})$.*

Proof. Given the definition of contribution function c, the lemma follows from Claim 1 and the assumption that Ψ satisfies property $\mathcal{R}_I$ (invariance). □

Claim 3. *$\beta_i{}^\epsilon - \beta_i = r$ for every $i \in C_1$, where $\beta = \Psi(M, \pi^b)$ and $\beta^\epsilon = \Psi(M^\epsilon, \pi^{b^\epsilon})$, and $r = \frac{1}{|C_1|} \cdot \sum_{S \subseteq \{1,\ldots,n\}} w \cdot [\Delta_S^\epsilon - \Delta_S]$.*

Proof. Notice that for every two agents $j, k \in C_1$ it holds that $c(M, \pi^b, j) = c(M, \pi^b, k)$ and that $\Delta_{S \cup \{j\}}^\epsilon - \Delta_{S \cup \{j\}} = \Delta_{S \cup \{k\}}^\epsilon - \Delta_{S \cup \{k\}} = \epsilon$ for every $S \subseteq \{1, ..., n\} \backslash \{j, k\}$. Furthermore,

from Claim 1 we have that $c(M, \pi^b, i) = c(M^\epsilon, \pi^{b^\epsilon}, i)$ for every i, and thus Lemma 6 applies, ${\beta_j}^\epsilon - \beta_j = {\beta_k}^\epsilon - \beta_k = r$, where r is some constant. Notice that:

$$r = \frac{1}{|C_1|} \cdot \sum_{i \in C_1} {\beta_i}^\epsilon - \beta_i.$$

By using Claim 2, we have that $r = \frac{1}{|C_1|} \cdot \sum_{i \in \{1,\dots,n\}} {\beta_i}^\epsilon - \beta_i$. Given that Ψ is assumed to satisfy $\mathcal{R}_{AE}$ (average efficiency), this implies that:

$$r = \frac{1}{|C_1|} \cdot \sum_{i \in \{1,\dots,n\}} {\beta_i}^\epsilon - \beta_i = \frac{1}{|C_1|} \cdot \sum_{S \subseteq \{1,\dots,n\}} w \cdot \left[\Delta_S^\epsilon - \Delta_S\right],$$

and hence ${\beta_i}^\epsilon - \beta_i = \frac{1}{|C_1|} \cdot \sum_{S \subseteq \{1,\dots,n\}} w \cdot [\Delta_S^\epsilon - \Delta_S]$ for every $i \in C_1$. □

Proofs of the Claims 4, 5, 6 and 7

Claim 4. *For each $\iota \in I$, $c(M, \pi^b, i) = c(M^\iota, \pi^{b^\iota}, i)$ for every i.*

Proof. Consider agent i such that $i \in C_1$. Given Eq. (9), this implies that $\Delta_i^\iota \geq \epsilon > 0 = \Delta_\emptyset^\iota$, and thus $c(M^\iota, \pi^{b^\iota}, i) = 1$. Hence, $c(M, \pi^b, i) = c(M^\iota, \pi^{b^\iota}, i)$.

Consider agent i such that $i \in C_0$. Given the definition of contribution function c, we have that $\Delta_{S \cup \{i\}} = \Delta_S$ for every S. By using Eq. (9), this implies that $\Delta_{S \cup \{i\}}^\iota = \epsilon = \Delta_S^\iota$ for every S such that $S \cap C_1 = S_\zeta$, where $\zeta > \iota$, $\Delta_{S \cup \{i\}}^\iota = \Delta_S^\iota$ for every S such that $S \cap C_1 = \emptyset$ and $\Delta_{S \cup \{i\}}^\iota = \Delta_{S \cup \{i\}} + \epsilon = \Delta_S + \epsilon = \Delta_S^\iota$ for every other S, and thus $c(M^\iota, \pi^{b^\iota}, i) = 0$. Hence, $c(M, \pi^b, i) = c(M^\iota, \pi^{b^\iota}, i)$. □

Claim 5. *For each $\iota \in I$, ${\beta_i}^\iota = 0$ for every $i \in C_0$, where $\beta^\iota = \Psi(M^\iota, \pi^{b^\iota})$.*

Proof. Given the definition of contribution function c, the lemma follows from Claim 4 and the assumption that Ψ satisfies property $\mathcal{R}_I$ (invariance). □

Based on the next observation we prove the rest of the claims:

Observation 1. *Observe that for each $\iota \in I \backslash \{2^{|C_1|} - 1\}$, $\Delta_S^{\iota+1} = \Delta_S^\iota$ for every S such that $S \cap C_1 \neq S_{\iota+1}$.*[11]

Claim 6. *For each $\iota \in I \backslash \{2^{|C_1|} - 1\}$, ${\beta_i}^{\iota+1} - {\beta_i}^\iota = 0$ for every $i \in C_1 \backslash S_{\iota+1}$, where $\beta^\iota = \Psi(M^\iota, \pi^{b^\iota})$ and $\beta^{\iota+1} = \Psi(M^{\iota+1}, \pi^{b^{\iota+1}})$.*

Proof. Notice that for every agent $i \in C_1 \backslash S_{\iota+1}$ it holds that $S \cup \{i\} \cap C_1 \neq S_{\iota+1}$ for every S. Given Observation 1 this implies that $\Delta_{S \cup \{i\}}^{\iota+1} = \Delta_{S \cup \{i\}}^\iota$ for every S. Furthermore, from Claim 4 we have that $c(M, \pi^b, i) = c(M^\iota, \pi^{b^\iota}, i) = c(M^{\iota+1}, \pi^{b^{\iota+1}}, i)$ for every i, and thus Lemma 5 applies, and for every $i \in C_1 \backslash S_{\iota+1}$ we have that ${\beta_i}^{\iota+1} = {\beta_i}^\iota$. □

Claim 7. *For each $\iota \in I \backslash \{2^{|C_1|} - 1\}$, ${\beta_i}^{\iota+1} - {\beta_i}^\iota = r$ for every $i \in S_{\iota+1}$, where $\beta^\iota = \Psi(M^\iota, \pi^{b^\iota})$ and $\beta^{\iota+1} = \Psi(M^{\iota+1}, \pi^{b^{\iota+1}})$, and $r = \frac{1}{|S_{\iota+1}|} \cdot \sum_{S \subseteq \{1,\dots,n\}} w \cdot \left[\Delta_S^{\iota+1} - \Delta_S^\iota\right]$.*

Proof. Notice that for every two agents $j, k \in S_{\iota+1}$ it holds that $c(M, \pi^b, j) = c(M, \pi^b, k)$. Given Claim 4, this implies that $c(M^\iota, \pi^{b^\iota}, j) = c(M^\iota, \pi^{b^\iota}, k)$. Notice also that $S \cup \{j\} \cap C_1 \neq S_{\iota+1}$ and $S \cup \{k\} \cap C_1 \neq S_{\iota+1}$ for every $S \subseteq \{1, ..., n\} \backslash \{j, k\}$. Given Observation 1, this implies that $\Delta_{S \cup \{j\}}^{\iota+1} - \Delta_{S \cup \{j\}}^\iota = \Delta_{S \cup \{k\}}^{\iota+1} - \Delta_{S \cup \{k\}}^\iota = 0$ for every $S \subseteq \{1, ..., n\} \backslash \{j, k\}$. Furthermore, from Claim 4 we have that $c(M, \pi^b, i) = c(M^\iota, \pi^{b^\iota}, i) = c(M^{\iota+1}, \pi^{b^{\iota+1}}, i)$ for every i, and thus Lemma

[11] Although it is not needed for the proofs of Claims 6 and 7, we mention that $\Delta_S^{\iota+1} = \Delta_S^\iota + \Delta_{S_{\iota+1}}$ for every S such that $S \cap C_1 = S_{\iota+1}$.

6 applies, and for every $j, k \in S_{\iota+1}$ we have that ${\beta_j}^{\iota+1} - {\beta_j}^{\iota} = {\beta_k}^{\iota+1} - {\beta_k}^{\iota} = r$, where r is some constant. Notice that:

$$r = \frac{1}{|S_{\iota+1}|} \cdot \sum_{i \in S_{\iota+1}} {\beta_i}^{\iota+1} - {\beta_i}^{\iota}.$$

By using Claim 5 and Claim 6, we have that $r = \frac{1}{|S_{\iota+1}|} \cdot \sum_{i \in \{1,\ldots,n\}} {\beta_i}^{\iota+1} - {\beta_i}^{\iota}$. Given that Ψ is assumed to satisfy $\mathcal{R}_{AE}$ (average efficiency), this implies that:

$$r = \frac{1}{|S_{\iota+1}|} \cdot \sum_{i \in \{1,\ldots,n\}} {\beta_i}^{\iota+1} - {\beta_i}^{\iota} = \frac{1}{|S_{\iota+1}|} \cdot \sum_{S \subseteq \{1,\ldots,n\}} w \cdot \left[\Delta_S^{\iota+1} - \Delta_S^{\iota}\right],$$

and hence ${\beta_i}^{\iota+1} - {\beta_i}^{\iota} = r = \frac{1}{|S_{\iota+1}|} \cdot \sum_{S \subseteq \{1,\ldots,n\}} w \cdot \left[\Delta_S^{\iota+1} - \Delta_S^{\iota}\right]$ for every $i \in S_{\iota+1}$. □

J Proofs of the Results from Section 4

This section of the appendix contains the proofs of the results from Section 3, in particular: Proposition 4, Proposition 5, and Theorem 3.

J.1 Proof of Proposition 4

Proposition 4. *Let $\widehat{\pi}^b$ be a solution to the optimization problem $\max_{\pi \in \mathcal{P}(\pi^b)} J(\pi)$. Then $\widehat{\Psi}_{SV,V}(M, \mathcal{P}(\pi^b)) = \Psi_{SV}(M, \widehat{\pi}^b)$ satisfies $\mathcal{R}_V$ (validity).*

Proof. In the setting of interest, $P(\pi^b)$ is consistent with π^b, that is $\pi^b \in P(\pi^b)$, and hence $J(\widehat{\pi}^b) \geq J(\pi^b)$. By Theorem 1, the blame attribution method Ψ_{SV} satisfies property $\mathcal{R}_E$ (efficiency), which implies that $\sum_{i=1}^{n} \widehat{\beta}_i = J(\pi^*) - J(\widehat{\pi}^b)$, where $\widehat{\beta} = \Psi_{SV}(M, \widehat{\pi}^b)$. This implies:

$$\begin{aligned}
&J(\widehat{\pi}^b) \geq J(\pi^b) \Rightarrow \\
\Rightarrow &J(\pi^*) - J(\widehat{\pi}^b) \leq J(\pi^*) - J(\pi^b) \Rightarrow \\
\Rightarrow &J(\pi^*) - J(\widehat{\pi}^b) \leq \Delta \Rightarrow \\
\Rightarrow &\sum_{i=1}^{n} \widehat{\beta}_i \leq \Delta.
\end{aligned}$$

Therefore, $\widehat{\Psi}_{SV,V}$ satisfies property $\mathcal{R}_V$ (validity). □

J.2 Proof of Proposition 5

Proposition 5. *Let β_i^i be the minimum value of the objective in* (P2)*. Then $\widehat{\Psi}_{SV,BC}(M, \mathcal{P}(\pi^b)) = (\beta_1^1, ..., \beta_n^n)$ satisfies $\mathcal{R}_V$ (validity) and $\mathcal{R}_{BC}(\Psi_{SV})$ (Blackstone consistency $w.r.t.$ $\Psi_{SV}(M, \pi^b)$).*

Proof. Let $\beta = \Psi_{SV}(M, \pi^b)$. Given Eq. (1), β_i^i being the minimum value of the objective in (P2) implies that $\beta_i^i = \min_{\pi \in \mathcal{P}(\pi^b)} {\beta_i}^{\pi}$ s.t. $\beta^{\pi} = \Psi_{SV}(M, \pi)$. In the setting of interest, $P(\pi^b)$ is consistent with π^b, that is $\pi^b \in P(\pi^b)$, which implies that $\beta_i^i = \min_{\pi \in \mathcal{P}(\pi^b)} {\beta_i}^{\pi} \leq \beta_i$. Therefore, $\widehat{\Psi}_{SV,BC}(M, \mathcal{P}(\pi^b))$ satisfies $\mathcal{R}_{BC}(\Psi_{SV})$ (Blackstone consistency $w.r.t.$ $\Psi_{SV}(M, \pi^b)$). Furthermore, by applying the same reasoning to all agents, we obtain $\sum_{i \in \{1,\ldots,n\}} {\beta_i}^{i} \leq \sum_{i \in \{1,\ldots,n\}} \beta_i$. Given Theorem 1, this implies $\sum_{i \in \{1,\ldots,n\}} {\beta_i}^{i} \leq \Delta$, and hence $\widehat{\Psi}_{SV,BC}(M, \mathcal{P}(\pi^b))$ also satisfies $\mathcal{R}_V$ (validity). □

J.3 Proof of Theorem 3

Theorem 3. *Consider $\widehat{\Psi}$ and Ψ s.t. $\left\| \widehat{\Psi}(M, \mathcal{P}(\pi^b)) - \Psi(M, \pi^b) \right\|_1 \leq \epsilon$ for any M, π^b, and $\mathcal{P}(\pi^b)$. Then if Ψ satisfies a property $\mathcal{R} \in \{\mathcal{R}_V, \mathcal{R}_E, \mathcal{R}_R, \mathcal{R}_S, \mathcal{R}_I, \mathcal{R}_{AE}\}$, $\widehat{\Psi}$ satisfies ϵ-$\mathcal{R}$. Moreover, if Ψ satisfies a property $\mathcal{R} \in \{\mathcal{R}_{CM}, \mathcal{R}_{PerM}, \mathcal{R}_{cPerM}, \mathcal{R}_{cParM}, \mathcal{R}_{RcParM}\}$, $\widehat{\Psi}$ satisfies 2ϵ-$\mathcal{R}$.*

Proof. We prove the implication for each property $\mathcal{R}$:

- $\mathcal{R}_V$ (validity): Let $\beta = \Psi(M, \pi^b)$ and $\widehat{\beta} = \widehat{\Psi}(M, \mathcal{P}(\pi^b))$. If Ψ satisfies $\mathcal{R}_V$,

$$\left\|\widehat{\beta} - \beta\right\|_1 \leq \epsilon \Rightarrow \sum_{i=1}^{n} |\widehat{\beta}_i - \beta_i| \leq \epsilon \Rightarrow \left|\sum_{i=1}^{n} \widehat{\beta}_i - \beta_i\right| \leq \epsilon \Rightarrow$$
$$\Rightarrow \sum_{i=1}^{n} \widehat{\beta}_i \leq \sum_{i=1}^{n} \beta_i + \epsilon \Rightarrow \sum_{i=1}^{n} \widehat{\beta}_i \leq \Delta + \epsilon,$$

  and hence $\widehat{\Psi}$ satisfies ϵ-$\mathcal{R}_V$.
- $\mathcal{R}_E$ (efficiency): Let $\beta = \Psi(M, \pi^b)$ and $\widehat{\beta} = \widehat{\Psi}(M, \mathcal{P}(\pi^b))$. If Ψ satisfies $\mathcal{R}_E$,

$$\left\|\widehat{\beta} - \beta\right\|_1 \leq \epsilon \Rightarrow \sum_{i=1}^{n} |\widehat{\beta}_i - \beta_i| \leq \epsilon \Rightarrow \left|\sum_{i=1}^{n} \widehat{\beta}_i - \beta_i\right| \leq \epsilon \Rightarrow$$
$$\Rightarrow \left|\sum_{i=1}^{n} \widehat{\beta}_i - \Delta\right| \leq \epsilon,$$

  and hence $\widehat{\Psi}$ satisfies ϵ-$\mathcal{R}_E$.
- $\mathcal{R}_R$ (rationality): Let $\beta = \Psi(M, \pi^b)$ and $\widehat{\beta} = \widehat{\Psi}(M, \mathcal{P}(\pi^b))$. If Ψ satisfies $\mathcal{R}_R$,

$$\left\|\widehat{\beta} - \beta\right\|_1 \leq \epsilon \Rightarrow \sum_{i=1}^{n} |\widehat{\beta}_i - \beta_i| \leq \epsilon \Rightarrow \sum_{i \in S} |\widehat{\beta}_i - \beta_i| \leq \epsilon \Rightarrow$$
$$\Rightarrow \left|\sum_{i \in S} \widehat{\beta}_i - \beta_i\right| \leq \epsilon \Rightarrow \sum_{i \in S} \widehat{\beta}_i \leq \sum_{i \in S} \beta_i + \epsilon \Rightarrow \sum_{i \in S} \widehat{\beta}_i \leq \Delta_S + \epsilon,$$

  and hence $\widehat{\Psi}$ satisfies ϵ-$\mathcal{R}_R$.
- $\mathcal{R}_S$ (symmetry): Let $\beta = \Psi(M, \pi^b)$ and $\widehat{\beta} = \widehat{\Psi}(M, \mathcal{P}(\pi^b))$. If Ψ satisfies $\mathcal{R}_S$,

$$\left\|\widehat{\beta} - \beta\right\|_1 \leq \epsilon \Rightarrow \sum_{i=1}^{n} |\widehat{\beta}_i - \beta_i| \leq \epsilon \Rightarrow |\widehat{\beta}_i - \beta_i| + |\widehat{\beta}_j - \beta_j| \leq \epsilon \Rightarrow$$
$$\Rightarrow \widehat{\beta}_i - \beta_i - \widehat{\beta}_j + \beta_j \leq \epsilon \Rightarrow \widehat{\beta}_i - \widehat{\beta}_j \leq \epsilon \tag{r1}$$

  and

$$\left\|\widehat{\beta} - \beta\right\|_1 \leq \epsilon \Rightarrow \sum_{i=1}^{n} |\widehat{\beta}_i - \beta_i| \leq \epsilon \Rightarrow |\widehat{\beta}_i - \beta_i| + |\widehat{\beta}_j - \beta_j| \leq \epsilon \Rightarrow$$
$$\Rightarrow -\widehat{\beta}_i + \beta_i + \widehat{\beta}_j - \beta_j \leq \epsilon \Rightarrow -\widehat{\beta}_i + \widehat{\beta}_j \leq \epsilon. \tag{r2}$$

  From (r1) and (r2), we have $|\widehat{\beta}_i - \widehat{\beta}_j| \leq \epsilon$, and hence $\widehat{\Psi}$ satisfies ϵ-$\mathcal{R}_S$.
- $\mathcal{R}_I$ (invariance): Let $\beta = \Psi(M, \pi^b)$ and $\widehat{\beta} = \widehat{\Psi}(M, \mathcal{P}(\pi^b))$. If Ψ satisfies $\mathcal{R}_I$,

$$\left\|\widehat{\beta} - \beta\right\|_1 \leq \epsilon \Rightarrow \sum_{i=1}^{n} |\widehat{\beta}_i - \beta_i| \leq \epsilon \Rightarrow |\widehat{\beta}_i - \beta_i| \leq \epsilon \Rightarrow$$
$$\Rightarrow |\widehat{\beta}_i| \leq \epsilon \Rightarrow \widehat{\beta}_i \leq \epsilon,$$

  and hence $\widehat{\Psi}$ satisfies ϵ-$\mathcal{R}_I$.
- $\mathcal{R}_{AE}$ (average efficiency): Let $\beta = \Psi(M, \pi^b)$ and $\widehat{\beta} = \widehat{\Psi}(M, \mathcal{P}(\pi^b))$. If Ψ satisfies $\mathcal{R}_{AE}$,

$$||\widehat{\beta} - \beta||_1 \leq \epsilon \Rightarrow \sum_{i=1}^{n} |\widehat{\beta}_i - \beta_i| \leq \epsilon \Rightarrow \left|\sum_{i=1}^{n} \widehat{\beta}_i - \beta_i\right| \leq \epsilon \Rightarrow$$
$$\Rightarrow \left|\sum_{i=1}^{n} \widehat{\beta}_i - \sum_{S \subseteq \{1,\ldots,n\}} \frac{1}{2^n - 1} \cdot \Delta_S\right| \leq \epsilon,$$

  and hence $\widehat{\Psi}$ satisfies ϵ-$\mathcal{R}_{AE}$.

- $\mathcal{R}_{CM}$ (contribution monotonicity) and $\mathcal{R}_{cParM}$ (c-participation monotonicity): Let $\beta^1 = \Psi(M^1, {\pi^b}^1)$, $\beta^2 = \Psi(M^2, {\pi^b}^2)$, $\widehat{\beta}^1 = \widehat{\Psi}(M^1, \mathcal{P}({\pi^b}^1))$ and $\widehat{\beta}^2 = \widehat{\Psi}(M^2, \mathcal{P}({\pi^b}^2))$. To show that Ψ satisfying $\mathcal{R}_{CM}$ (resp. $\mathcal{R}_{cParM}$) implies that $\widehat{\Psi}$ satisfies 2ϵ-$\mathcal{R}_{CM}$ (resp. 2ϵ-$\mathcal{R}_{cParM}$), it suffices to show that ${\beta_i}^1 - {\beta_i}^2 \geq 0$ implies ${\widehat{\beta}_i}^1 \geq {\widehat{\beta}_i}^2 - 2\epsilon$. Let ${\beta_i}^1 - {\beta_i}^2 \geq 0$. We have

$$\left\|\widehat{\beta}^1 - \beta^1\right\|_1 \leq \epsilon \Rightarrow \sum_{i=1}^{n} |{\widehat{\beta}_i}^1 - {\beta_i}^1| \leq \epsilon \Rightarrow |{\widehat{\beta}_i}^1 - {\beta_i}^1| \leq \epsilon \Rightarrow$$

$$\Rightarrow {\beta_i}^1 - {\widehat{\beta}_i}^1 \leq \epsilon \tag{r3}$$

and

$$\left\|\widehat{\beta}^2 - \beta^2\right\|_1 \leq \epsilon \Rightarrow \sum_{i=1}^{n} |{\widehat{\beta}_i}^2 - {\beta_i}^2| \leq \epsilon \Rightarrow |{\widehat{\beta}_i}^2 - {\beta_i}^2| \leq \epsilon \Rightarrow$$

$$\Rightarrow {\widehat{\beta}_i}^2 - {\beta_i}^2 \leq \epsilon. \tag{r4}$$

By adding (r3) and (r4), we obtain

$${\beta_i}^1 - {\widehat{\beta}_i}^1 + {\widehat{\beta}_i}^2 - {\beta_i}^2 \leq 2\epsilon \Rightarrow {\widehat{\beta}_i}^1 \geq {\widehat{\beta}_i}^2 - 2\epsilon,$$

and hence $\widehat{\Psi}$ satisfies 2ϵ-$\mathcal{R}_{CM}$ (resp. 2ϵ-$\mathcal{R}_{cParM}$).

- $\mathcal{R}_{PerM}$ (performance monotonicity) and $\mathcal{R}_{cPerM}$ (c-performance monotonicity): Let $\beta = \Psi(M, (\pi_i, {\pi^b}_{-i}))$, $\beta' = \Psi(M, (\pi'_i, {\pi^b}_{-i}))$, $\widehat{\beta} = \widehat{\Psi}(M, \mathcal{P}((\pi_i, {\pi^b}_{-i})))$ and $\widehat{\beta}' = \widehat{\Psi}(M, \mathcal{P}((\pi'_i, {\pi^b}_{-i})))$. To show that Ψ satisfying $\mathcal{R}_{PerM}$ (resp. $\mathcal{R}_{cPerM}$) implies that $\widehat{\Psi}$ satisfies 2ϵ-$\mathcal{R}_{PerM}$ (resp. 2ϵ-$\mathcal{R}_{cPerM}$), it suffices to show that $\beta_i \geq {\beta_i}'$ implies $\widehat{\beta}_i \geq {\widehat{\beta}_i}' - 2\epsilon$. Let $\beta_i \geq {\beta_i}'$. We have

$$\left\|\widehat{\beta} - \beta\right\|_1 \leq \epsilon \Rightarrow \sum_{i=1}^{n} |\widehat{\beta}_i - \beta_i| \leq \epsilon \Rightarrow |\widehat{\beta}_i - \beta_i| \leq \epsilon \Rightarrow$$

$$\Rightarrow \beta_i - \widehat{\beta}_i \leq \epsilon \tag{r5}$$

and

$$\left\|\widehat{\beta}' - \beta'\right\|_1 \leq \epsilon \Rightarrow \sum_{i=1}^{n} |{\widehat{\beta}_i}' - {\beta_i}'| \leq \epsilon \Rightarrow |{\widehat{\beta}_i}' - {\beta_i}'| \leq \epsilon \Rightarrow$$

$$\Rightarrow {\widehat{\beta}_i}' - {\beta_i}' \leq \epsilon. \tag{r6}$$

By adding (r5) and (r6), we obtain

$$\beta_i - \widehat{\beta}_i + {\widehat{\beta}_i}' - {\beta_i}' \leq 2\epsilon \Rightarrow \widehat{\beta}_i \geq {\widehat{\beta}_i}' - 2\epsilon,$$

and hence $\widehat{\Psi}$ satisfies 2ϵ-$\mathcal{R}_{PerM}$ (resp. 2ϵ-$\mathcal{R}_{cPerM}$).

- $\mathcal{R}_{RcParM}$ (relative c-participation monotonicity): Let $\beta^1 = \Psi(M^1, {\pi^b}^1)$, $\beta^2 = \Psi(M^2, {\pi^b}^2)$, $\widehat{\beta}^1 = \widehat{\Psi}(M^1, \mathcal{P}({\pi^b}^1))$ and $\widehat{\beta}^2 = \widehat{\Psi}(M^2, \mathcal{P}({\pi^b}^2))$. To show that Ψ satisfying $\mathcal{R}_{RcParM}$ implies that $\widehat{\Psi}$ satisfies 2ϵ-$\mathcal{R}_{RcParM}$, it suffices to show that ${\beta_j}^1 - {\beta_j}^2 \geq {\beta_k}^1 - {\beta_k}^2$ implies ${\widehat{\beta}_j}^1 - {\widehat{\beta}_j}^2 \geq {\widehat{\beta_k}}^1 - {\widehat{\beta_k}}^2 - 2\epsilon$. Let ${\beta_j}^1 - {\beta_j}^2 \geq {\beta_k}^1 - {\beta_k}^2$. We have

$$\left\|\widehat{\beta}^1 - \beta^1\right\|_1 \leq \epsilon \Rightarrow \sum_{i=1}^{n} |{\widehat{\beta}_i}^1 - {\beta_i}^1| \leq \epsilon \Rightarrow |{\widehat{\beta}_j}^1 - {\beta_j}^1| + |{\widehat{\beta_k}}^1 - {\beta_k}^1| \leq \epsilon \Rightarrow$$

$$\Rightarrow {\beta_j}^1 - {\widehat{\beta}_j}^1 - {\beta_k}^1 + {\widehat{\beta_k}}^1 \leq \epsilon \tag{r7}$$

and

$$\left\|\widehat{\beta}^2 - \beta^2\right\|_1 \leq \epsilon \Rightarrow \sum_{i=1}^{n} |{\widehat{\beta}_i}^2 - {\beta_i}^2| \leq \epsilon \Rightarrow |{\widehat{\beta}_j}^2 - {\beta_j}^2| + |{\widehat{\beta_k}}^2 - {\beta_k}^2| \leq \epsilon \Rightarrow$$

$$\Rightarrow -\beta_j{}^2 + \widehat{\beta}_j{}^2 + \beta_k{}^2 - \widehat{\beta_k}^2 \leq \epsilon. \tag{r8}$$

By adding (r7) and (r8), we obtain

$$\begin{aligned} &\beta_j{}^1 - \widehat{\beta}_j{}^1 - \beta_k{}^1 + \widehat{\beta_k}^1 - \beta_j{}^2 + \widehat{\beta}_j{}^2 + \beta_k{}^2 - \widehat{\beta_k}^2 \leq 2\epsilon \Rightarrow \\ \Rightarrow &\widehat{\beta}_j{}^1 - \widehat{\beta}_j{}^2 \geq \widehat{\beta_k}^1 - \widehat{\beta_k}^2 - 2\epsilon, \end{aligned}$$

and hence $\widehat{\Psi}$ satisfies 2ϵ-$\mathcal{R}_{RcParM}$.

□