Measuring Cooperation with Counterfactual Planning

Samuel A. Barnett[1[0000−0001−9612−3096]], Kathryn Wantlin[1[0009−0006−1430−8729]], and Ryan P. Adams[1[0000−0002−5704−6654]]

Department of Computer Science, Princeton University, Princeton NJ 08544, USA
samuelab@alumni.princeton.edu, {kw2960, rpa}@princeton.edu

Abstract. Cooperative behavior is commonly understood as that which is conducive to the good of the group: it is increasingly seen as a crucial component of advancing the capabilities as well as mitigating the harms of multi-agent AI systems [21, 6, 10]. Yet an "I'll-know-it-when-I-see-it" approach is often taken when evaluating the cooperativeness of a sequence of actions, and even when cooperation is formalized, the definitions lead to category errors, conceptual confusions, and erroneous conclusions [22, 11, 56, 52]. We propose a formal measure of cooperation in stochastic games that avoids these pitfalls by being *counterfactually contrastive*, *contextual*, and *customizable*: in particular, cooperation is defined in contrast to the outcome that a self-interested actor would have effected in a similar circumstance, in the context of other agents' behavior, and within a specified time and space horizon. This measure is simple to compute: it is dependent only on solving a reduction of the multi-agent game to a single-agent Markov decision process. We apply this measure to a diverse pool of behaviors in a number of mixed-motive social dilemmas and sequential predator-prey environments that have been studied in the multi-agent systems literature [4, 26, 34, 15, 36]. Our results demonstrate the importance of defining cooperation clearly, and provide a useful metric for builders of cooperative systems to use when establishing the cooperative nature of the system behavior.

Keywords: Cooperation · Multi-Agent Systems · Reinforcement Learning · Social Dilemmas.

1 Introduction

In the trees of the Tai Forest in Côte d'Ivoire, chimpanzees hunt for red colobus monkeys in groups. Each chimpanzee shares the goal of hunting the monkey, and each chimpanzee benefits from the participation of the other chimpanzees in order to increase the likelihood that the prey is caught. Therefore, each chimpanzee is acting in a way that is conducive to the good of the group—this would appear to be a paradigmatic case of cooperative behavior.

However, there is another characterization of this sequence of events [47]. One chimpanzee initiates the hunt in the knowledge that other chimpanzees are in the area, and then each other chimpanzee will in turn take the position that best

maximizes its own likelihood of catching the prey. This has the cumulative effect of each chimpanzee blocking the monkey's next best path of escape. Importantly, each of the chimpanzees takes these actions individually and makes its plans solely according to its own self-regard; there is no central planning.

A similar dynamic can arise within artificial systems. Although central planning is possible in theory and may be more likely to lead to desirable outcomes, due to its computational demand the approach is often eschewed in favor of agents who learn and act *independently* in an environment without regard to the other agents' utilities [11]. In many cases, this can still lead to an outcome that is beneficial to all of the agents [44].

In order to evaluate the cooperativeness of group behavior in both artificial multi-agent systems and biological species, we need to be able to measure the cooperativeness of these systems [7, 6]. However, as the preceding examples show, behavior that increases the total utility of the group is not necessarily cooperative—in other words, the cooperativeness of behavior is *underdetermined* by the actual sequence of events [31].

Previous work studying cooperation in artificial systems has focused on the design of environments within which cooperation can be understood, using these to investigate what mechanisms can drive cooperation [4, 22, 15, 16, 32, 10, 9]. However, cooperative behavior is typically either declared so by fiat, or is defined only in relation to the actual group outcome, without reference to the actions of uncooperative agents. An alternative to this is to measure the alignment between individual and collective interests in the system as a whole, such as through the *price of anarchy* [19] or the *self-interest level* [53].

In this paper, we propose a family of scalar measures of cooperation capable of precluding cases such as that of unintended mutual benefit by being *counterfactually contrastive*: we subtract from the group's total utility the amount that would have been attained had the agent in question acted purely in their self-interest. Our approach is agnostic to the mechanisms that distinguish between cooperative and competitive modes of group behavior [17, 46, 45], and it does not require any manipulations of the external rewards in the environment [27]. Moreover, we allow our measure to be *contextual*, in that it is relative to other agents' behavior, as well as *customizable* with respect to the time and space horizons, which can help to disambiguate other gray areas of cooperative behavior that have been previously studied.

We define our measure on stochastic games, a formalization of multi-agent systems that allow for the application of our measure on a broad class of artificial agents, as well as biological agents that can be modelled in this way [41]. Using this definition, we evaluate the behavior of multiple classes of agents with different types of behavior in tabular social dilemmas, a common test bed in a variety of disciplines for understanding cooperation [4], as well as more complex predator-prey environments. We show that many of these behaviors are no longer regarded as cooperative when our measure is applied to it, and other seemingly uncooperative behaviors become otherwise according to our measure. Crucially, by making explicit the components of the measure, our measure can provide

an interpretable explanation of why behavior is cooperative or uncooperative. Moreover, by making the choice of social welfare function one of these components, our measure also explains the respect in which this behavior is cooperative, either by achieving a greater total utility, or a more equitable or fair outcome.

2 Related work

Cooperation has long been a subject of study in disciplines ranging from philosophy and economics to evolutionary biology and cognitive science [4, 47, 49, 20]. For a comprehensive review of the study of cooperation in the context of multi-agent RL, refer to [55].

In order to study cooperation with computational models, an initial approach is simply to declare behaviors as cooperative or defecting by fiat. For example, in the Prisoner's Dilemma, the classic one-shot social dilemma game, the available actions to each player are to "Cooperate" or "Defect". The conclusions drawn from analyses of this game are subsequently generalized about cooperation as a broader concept [4]. While there are some attempts to expand this approach to multi-step environments by measuring the success of introduced mechanisms or the frequency that certain tasks are performed,[1] this approach generally fails as we begin to examine systems acting in more complex environments that are capable of a richer range of behaviors. In particular, these behaviors will arguably now be cooperative to different *degrees*, with the cooperativeness of each behavior not necessarily being obvious [22].

Hence the need to define a *measure* on the cooperativeness of behavior. At a first pass, we might do this by simply evaluating the sum of all utilities attained by the group, also referred to as the *utilitarian welfare* [15, 23]; other welfare metrics such as fairness or sustainability could also be considered [3]. One drawback of this approach is that the cooperativeness of behavior is defined on groups as a whole, whereas it would be desirable for a measure to tell us if one agent were acting *more* cooperatively than others within the group.

More importantly, however, solely evaluating the actual outcome erroneously includes cases such as the aforementioned chimpanzee group hunting in which a mutually beneficial outcome results from individual agents acting solely in their own self-regard. A related phenomenon occurs in evolutionary biology in which two species feed upon the waste product of the other: this is known as byproduct reciprocity. Unless this behavior is selected for *because* of the beneficial effect on the recipient (or at least partially because of this effect), this is not classed as cooperation [52].

Another issue when defining cooperation relates to the time horizon over which it is evaluated. The utility accrued from a group behavior, either to the individual or the entire group, may vary in its magnitude and valence over time. For example, *reciprocally altruistic* individuals take turns helping each other in a costly way with the expectation that they will be helped in the future [48].

[1] Refer to [9] for an overview of such metrics in the cooperative multi-agent learning literature.

The term "altruism" is commonly taken to be a misnomer in relation to this phenomenon [13, 52], and this mistake can be clarified with appeal to the time horizon in question: while the behavior is costly to the agent performing it in the short term, in the long term we expect that the reciprocated benefits will justify this cost, so that the behavior can eventually be considered as self-interested.

3 Desiderata for a measure of cooperation

To address these common pitfalls, we divide the desiderata for a measure of cooperation into three broad categories: that it should be *counterfactually contrastive*, *customizable*, and *contextual*.

Counterfactually contrastive The absolute returns in total utility can be a misleading guide to the cooperativeness of a system: in certain situations, these returns might result without any cooperation taking place. We call a measure *counterfactually contrastive* if it sets as a baseline the behavior of agent(s) acting uncooperatively.

We base our measure on the contrast between an agent acting in accordance with its own goals, rather than the goals of the group. This is much simpler to evaluate for any given agent, as we need only consider what that agent's best response is to the behavior of the rest of the system, assuming the agent is only concerned with its own goals. This captures the individualized description of what truly occurs during chimpanzee group hunting [47].

Customizable A measure of cooperation should be *customizable* insofar as it allows variations to certain components that are key to determining how cooperative a given behavior is.

One such component is the *time scale* over which the behavior is evaluated. This is important for understanding the challenge posed by direct fitness explanations of cooperation: the self-interested benefits of direct fitness accrue only in the long term, whereas in the short term the behavior in question may seem counterintuitive from the self-interested perspective. Moreover, this addresses the challenge posed by cases of reciprocal altruism, as discussed in the previous section.

Another important component of any evaluation of cooperative behavior is the way in which social outcomes are valued: while a typical choice would be to take the sum of all relevant agents' utilities, this does not always capture everything that we care about for a given outcome. For instance, we might instead evaluate success in terms of the utility of the worst-off agent, or in terms of the equitability of outcomes for each individual agent. The choice of metric for each social outcome will vary depending on the multi-agent system in question, and should not be held as a fixed component of the cooperative measure.

Contextual Finally, a measure should be *contextual*, so as to reflect the idea that the cooperativeness of an individual agent's actions depends on those of the other agents in the system they are interacting in. Hence, when measuring

the cooperativeness of the agent, it will always be relative to the other agents in question.

By making the measure contextual, we also make explicit the subgroup of agents on which the social outcomes are considered. While this subgroup may include all of the agents in the environment, this is not a requirement: in the example of predator-prey interactions, we do not consider the utility of the prey to be a factor in the cooperativeness of the group hunting behavior.

4 Measuring cooperation in stochastic games

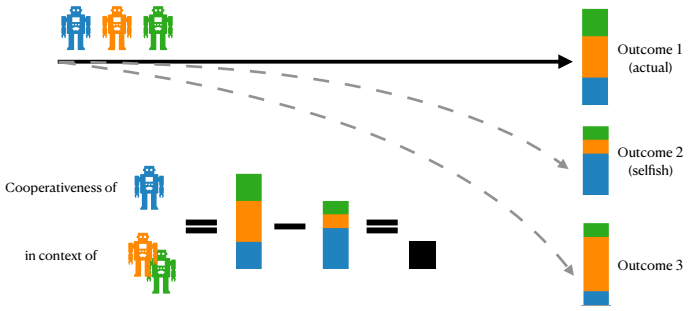

Fig. 1. A schematic of the cooperation measure. A multiagent system consisting of three robots (Blue, Orange, and Green) has three outcomes, one actual and two potential, with utilities to each agent represented by stacked colored bars. The cooperativeness measure for Blue's policy consists of subtracting from the actual welfare (given here by total utility) the welfare for the selfish outcome, that is, the outcome for which Blue's utility is the largest.

We define our measure of cooperation within the framework of *stochastic games* [40, 41], defined as a tuple (S, N, A, P, R, γ) consisting of a state space S, a finite set of agents N indexed by i, a set of available actions $A = \prod_i A^{(i)}$ for each agent, a transition probability function $P\colon S \times A \times S \to [0, 1]$, a scalar reward function for each agent $R = (r^{(1)}, ..., r^{(N)})$, where $r^{(i)}\colon S \times A \to \mathbb{R}$, and a discount factor $\gamma \in [0, 1]$.

Each state s can be considered as its own normal form game. We can therefore consider the space of Markov strategies (or policies) $\pi^{(i)}\colon S \times A^{(i)} \to [0, 1]$ specifying the probability of taking each possible action $a^{(i)}$ in state s. Assuming the agents follow policies $\boldsymbol{\pi} = (\pi^{(1)}, ..., \pi^{(N)})$, we define the *value* of a state s for agent i to be the expected discounted sum of rewards for that agent:

$$V_{\boldsymbol{\pi}}^{(i)}(s) = \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=T}^{\infty} \gamma^{t-T} r^{(i)}(s_t, \boldsymbol{a}_t) \,\middle|\, s_T = s\right]. \tag{1}$$

To measure the social outcome of a stochastic game, we use a welfare metric w that is a function of each agent's value function and some distribution over the states $\rho \in \Delta S$. A typical choice for these is to use the *utilitarian* welfare weighted by the initial state distribution as our metric, $w_U : (\boldsymbol{\pi}; \rho_0) \mapsto \sum_{s \in S} \sum_{i=1}^{N} \rho_0(s) V_{\boldsymbol{\pi}}^{(i)}(s)$.

If we fix the policies of all agents except for i (denoting these as $\boldsymbol{\pi}^{(-i)}$), the stochastic game reduces to a single-agent Markov Decision Process (MDP). Let $\mathrm{BR}_i\left(\boldsymbol{\pi}^{(-i)}\right)$ denote the (non-empty) set of optimal policies (or *best responses*) for agent i in the context of the other agents choosing policies $\boldsymbol{\pi}^{(-i)}$, i.e., the set of solutions to the single-agent MDP.

Finally, we define our measure of cooperation for a policy $\pi^{(i)}$ in the context of $\boldsymbol{\pi}^{(-i)}$ as the welfare of these policies, minus the best possible welfare of agent i's best response policy:

$$c\left(\pi^{(i)}; \boldsymbol{\pi}^{(-i)}\right) = w\left(\pi^{(i)}, \boldsymbol{\pi}^{(-i)}\right) - \max_{\pi_*^{(i)} \in \mathrm{BR}_i\left(\boldsymbol{\pi}^{(-i)}\right)} w\left(\pi_*^{(i)}, \boldsymbol{\pi}^{(-i)}\right). \tag{2}$$

A schematic diagram explaining this definition can be seen in Fig. 1.

By defining a scalar measure for cooperation, we are now able to evaluate the *degree* to which a policy is cooperative or uncooperative, and therefore we can also make comparative judgements between different policies. Intuitively, the measure evaluates the extent to which policy $\pi^{(i)}$ improves the social welfare over the (best) outcome that would have resulted from the agent acting purely in its self-interest.

This definition is clearly contextual, as the cooperativeness of $\pi^{(i)}$ depends on the context of the other agents' policies $\boldsymbol{\pi}^{(-i)}$. Moreover, the definition clearly depends on the choice of welfare and discount factor. Each of these serves as an example of the measure's customizability, with the discount factor capturing the notion of a relevant time horizon.

The measure is also counterfactually contrastive in the sense set out above. In this case, we take the relevant counterfactual to hold fixed the policies of the other agents, and consider a self-interested agent to be one who maximizes its value in response to these policies. This contrasts with the counterfactuals considered by cooperative MARL algorithms such as COMA [12] and SHAQ [51], in which only the *actions* of other agents are held fixed.

It further contrasts with credit assignment methods from cooperative game theory such as the *Shapley value*, which poses a different kind of counterfactual based on an agent's presence or absence from a group [39]. While the Shapley value answers a combinatorial question about the value of an agent's participation, our approach instead considers the continuous value of policies to evaluate how an agent acted relative to how it might have acted. This allows for a more nuanced analysis of behavior within a fixed set of agents, a question that the Shapley value does not address.

In environments represented by a sufficiently small state space, we can compute optimal policies to arbitrary precision with value iteration [43], although for more complex systems we can also approximate the cooperation measure by using reinforcement learning to find approximate solutions to this problem.

5 Experiments

5.1 Matrix Game Social Dilemmas

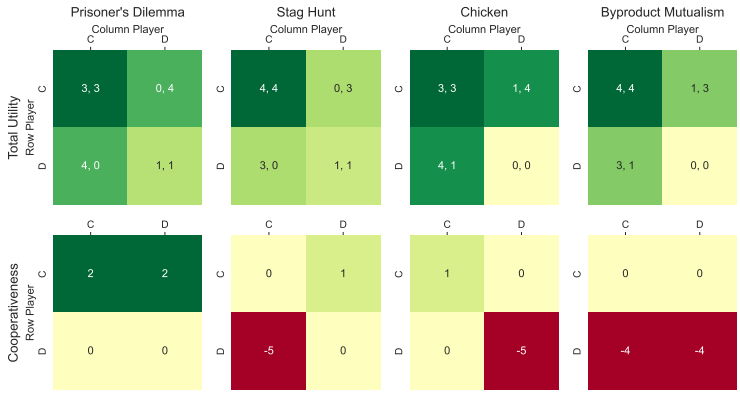

Fig. 2. The four varieties of matrix game social dilemmas: the Prisoner's Dilemma, Stag Hunt, Chicken, and Byproduct Mutualism. The top row shows the payoff matrices for each game, with the colors representing the value of the total utility (calculated by adding the players' payoffs in each cell). The bottom row shows the heatmaps for the cooperativeness score for the row player's action in the context of the column player's action.

To motivate the applicability of our measure, we begin by evaluating the cooperativeness of different strategies in four matrix games. The first three of these are the canonical *one-shot social dilemmas* that are designed to elucidate the opposing pressures of individual rationality and ideal collective action [8, 25, 35, 26]. These dilemmas are therefore designed so as to clearly differentiate between cooperative and uncooperative behavior in a way that ought to be apparent in our measure.

In these games, two agents have the choice of actions C (for *Cooperate*) or D (for *Defect*). The agents prefer mutual C to mutual D, mutual C to unilateral C, and mutual C yields a higher total utility than mutual D. However, in each game, we have that either unilateral D is preferable to mutual C (so that you can do better by exploiting a cooperator than cooperating with one), or that mutual D is preferable to unilateral C (so that being exploited is worse than not cooperating with a would-be exploiter). Chicken meets only the first of these disjuncts, Stag Hunt only the second, and Prisoner's Dilemma meets both.

The bottom row of Fig. 2 shows the cooperativeness of each row player's action in the context of the column player's action, using total utility as welfare. Holding fixed this context, we see that C is always strictly more cooperative

than D, supporting the interpretation of C and D as *cooperation* and *defection*, respectively. Notably, in Byproduct Mutualism there is no pair of actions with a positive cooperativeness score. This is due to the fact that in this game the dilemma is completely relaxed: it is better to cooperate irrespective of the partner's decision, and so the choices that lead to the highest collective utility are also precisely the ones that self-interested actors would take.

5.2 Iterated Social Dilemmas

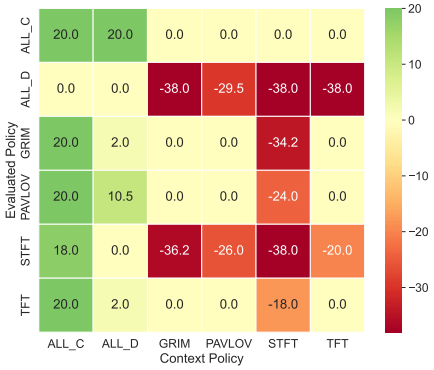

Fig. 3. Cooperativeness of six common deterministic policies in the Iterated Prisoner's Dilemma, in the context of each other policy. The cooperativeness is valued on the initial state, with a discount factor $\gamma = 0.9$.

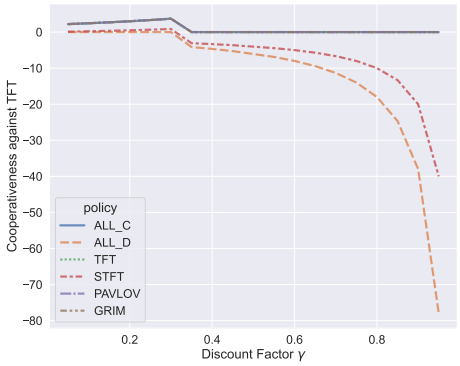

Fig. 4. Cooperativeness of six common deterministic policies in the context of Tit-for-Tat in the iterated Prisoner's Dilemma, plotted against the discount factor γ. As the `ALL_C`, `TFT`, `PAVLOV`, and `GRIM` strategies all cooperate on the initial time-step, their outcomes playing against `TFT` are identical and so their cooperativeness ratings overlap.

When we move to the iterated Prisoner's Dilemma, in which agents interact in a Prisoner's Dilemma *ad infinitum*, there is no strictly dominant individual strategy in this game.[2] Nonetheless, a number of strategies have been proposed with desirable properties [4, 30, 42]. We limit our strategies to those that depend on at most one previous interaction, referred to as *memory-1 strategies*: this includes strategies such as (Suspicious)-Tit-for-Tat (`(S)TFT`), Win-Stay-Lose-Shift (`PAVLOV`), and Grim (`GRIM`) (and their stochastic variants), but excludes others such as Tit-for-Two-Tats or Majority that require keeping track of a longer history. Hence, the MDP that arises from fixing the opponent's strategy to one of these will have five states (one for each possible action combination, and an additional initial state), making it tractable to solve so that we can compute the cooperativeness scores to arbitrary precision with value iteration [43].

[2] Refer to Appendix B for analyses of the iterated Chicken and Stag Hunt.

Fig. 3 shows the cooperativeness measure applied to six deterministic memory-1 strategies, with each strategy being evaluated in the context of the other agent adopting every other strategy from the group. Strategies that take action C in more states generally score higher than strategies that take action D. However, we also see that the context policy plays an important role in determining the cooperativeness of the evaluated policy. In particular, in the context of policies that punish defection (either for one turn as in the case of `(S)TFT` or forever as in the case of `GRIM`), `ALL_C` does not rank as cooperative, as it becomes the best-response strategy. This supports the intuition that cooperating in the face of potential punishment is not as cooperative as unconditional cooperation, allowing us to distinguish between coercion and cooperation [37].

We also see the impact that the discount factor has on the measure of cooperativeness. If the column player adopts the `TFT` policy, then a row player will be able to exploit the fact that this strategy cooperates in the initial turn, at the expense of a defection in the subsequent turn. Therefore, if future rewards are sufficiently discounted relative to immediate rewards, it is optimal for the row player to initially defect. However, if future rewards are not significantly discounted, then it is in the row player's best interest to always cooperate. Fig. 4 shows the cooperativeness of each memory-1 strategy in the context of `TFT` plotted against the discount factor: cooperative policies such as `ALL_C` score positively on cooperativeness for lower values of the discount factor, with the score eventually tending towards zero. On the other hand, defecting policies such as `ALL_D` have cooperativeness scores that begin at zero before tending towards $-\infty$.

5.3 Tabular Cleanup

Though iterated matrix games can lead to a richer range of behaviors through the use of memory-based strategies, the actions themselves that these strategies are defined over nonetheless treat *cooperate* and *defect* as primitives. A more faithful depiction of social dilemmas demands more complex strategies that apply to *policies* over richer action and state space. To this end, we investigate a simplified version of the social dilemma *Cleanup* [34, 15, 1, 14, 50]. This is an example of a public goods dilemma, in which an individual must pay a personal cost in order to provide a resource that is shared by all [18].

This game consists of N players who can choose between the actions *Clean*, *Eat*, and *Punish Player* i for $i = 1, ..., N$. The state space consists of the actions taken by each player at the previous time-step, and the number of apples currently available, which can range from 0 to $3N - 1$. An apple grows with a probability linearly proportional to the number of agents choosing *Clean*, with the probability ranging from 0 to 1. If an agent chooses to eat an apple, it receives a reward of $+1.0$, unless there are fewer apples available than agents eating, in which case the reward is divided amongst the eaters. If an agent chooses to punish another agent, it imposes a -2.0 reward deduction from the target, at an expense of -0.5 reward. For $N = 2$, we exclude the possibility of self-punishment for simplicity.

We consider two- and three-player instantiations of *Tabular Cleanup*, leading to state spaces of sizes 54 and 1125, respectively. This includes states which are

not reachable from any other state: we therefore define the distribution over initial states according to the states reached by starting with a uniformly random choice over actions and numbers of apples. We evaluate the following policies:[3]

- *Always* X: This policy takes a constant action across all states.
- *Take Turns*: This policy alternates between cleaning and eating.
- *TFT*: This policy reciprocates the action taken by its co-player in the previous timestep.
- *Nash*: This policy cleans when there are no apples available, and eats otherwise. As the name suggests, this is a Nash equilibrium to the game.
- *Prosocial*: This policy cleans when there are fewer apples available than the number of players, and eats otherwise. This was derived by solving the MDP derived from the two-player game with a centralized actor controlling both agents, and a reward consisting of the sum of the player's rewards.

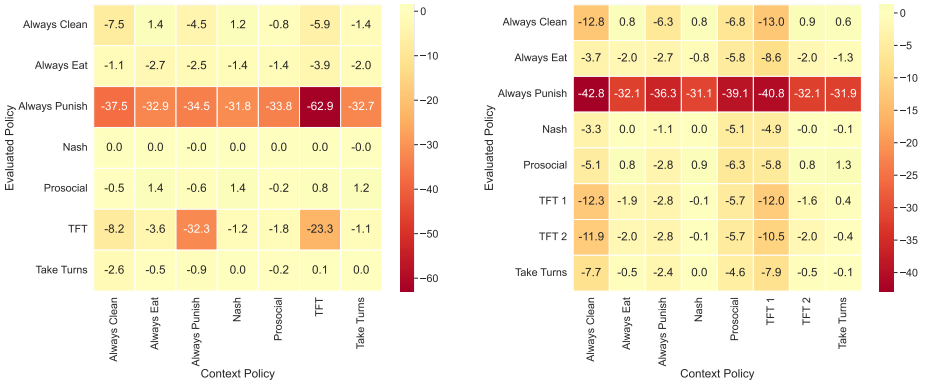

Fig. 5. Cooperativeness of deterministic policies in the Tabular Cleanup game, evaluated on the initial state with a discount factor of $\gamma = 0.9$. **Left:** Seven policies in the 2-player version, each in the context of every other policy. **Right:** Eight policies in the 3-player version, where the acting policy is in the context of two other players using the same policy.

Fig. 5 shows the results of evaluating each of these policies in a variety of contexts. As expected, the *Prosocial* policy is the most cooperative on average across all contexts, and *Always Punish* the least. However, in many contexts, *Always Clean* is less cooperative than *Always Eat*, and *Take Turns* is more cooperative than both. This can be explained by the fact that eating contributes to the collective reward through adding to your *own* reward, and so choosing to clean in states where there are a sufficient number of apples for all agents needlessly forgoes a reward that contributes to the joint welfare. While such a

[3] Refer to Sec A for the definitions of the three-player Tabular Cleanup policy variants.

result is intuitive, it is obscured by discussions of *Cleanup* that simply equate cooperativeness with the frequency at which each agent cleans [15].

These results can also be interpreted as providing a quantitative argument that *specialization* can be crucial to cooperation depending on the context. In the context of an agent that always eats, it is in fact more cooperative to focus on cleaning. However, in the converse context, eating becomes more imperative for increasing the joint welfare.

5.4 Multi-agent RL Environments

Finally, we evaluate our measure in partially-observable stochastic games, in which each agent has only incomplete information on the state of the game by limiting each agent's field of vision to a small subgrid of pixel values. Training agents to maximize rewards in such games typically requires deep reinforcement learning algorithms such as PPO [38, 54], in addition to policy models based on neural networks that must first learn to map pixel observations onto appropriate features. Despite these challenges, our measure of cooperativeness is sufficiently general to capture such cases by finding an "approximate best-response" that gives a cooperativeness upper-bound.

Due to this increased complexity, we are no longer able to straightforwardly define policies by specifying actions on individual states. Instead, we define different policies by changing the conditions of the environment in which the RL algorithm learns a policy, interpreting each policy in relation to the conditions under which it is trained [22]. In particular, we refer to a policy as *selfish* if it is trained to maximize the individual value of the agent following it, and *prosocial* if it is trained to maximize the sum of the values of *all* agents. The goal of this experiment is to investigate whether these naïve interpretations indeed align with the cooperativeness scores attained for the policies.

We investigate the Simple Tag game, a multi-particle environment [29] in which "predator" agents pursue "prey" agents for reward. While the reward structure is typically implemented so that all predators share the reward of catching individual prey, we contrast this prosocial version with the selfish version of predators only receiving rewards for the prey that they have caught. The prey in this case follows a heuristic policy, as in [33].

This game forms a part of JaxMARL suite of benchmarks [36]. By writing the policy models, the PPO algorithm, *and* the environments themselves exclusively in JAX [5], this suite can leverage GPU acceleration, automatic vectorization, and just-in-time XLA compilation to implement a training pipeline with orders of magnitude greater efficiency than that of its counterparts. In particular, we are able to run our experiments over 160 random seeds in a matter of minutes.

Fig. 6 shows the results for Simple Tag with 2 predators and 1 heuristic prey. We observe several pieces of evidence that in this case the "selfish" policy is more cooperative than the "prosocial" policy: (i) it has a greater cooperativeness score in both contexts, with the difference being larger in the context of a "selfish" policy; (ii) the greater the number of "selfish" policies, the greater the actual welfare achieved; and (iii) a "selfish" policy in the context of a "prosocial" policy

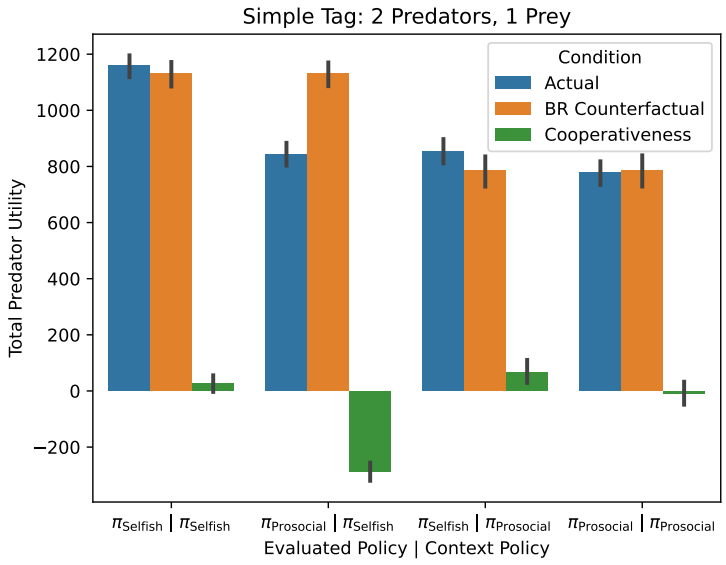

Fig. 6. Results for Simple Tag with 2 predators (and one prey). The y-axis shows the total utility for the predator under three conditions: the actual outcome for the evaluated and context policy, the counterfactual of the first agent pursuing a selfish best-response against the fixed context policy, and the cooperativeness measuring the difference between the two. These results are taken over 160 seeds, with error bars showing the 95% CI.

is positively cooperative, and the opposite is anticooperative. However, as in the hunting scenario described in the introduction, a "selfish" policy in the context of other "selfish" policies yields a cooperativeness score that does not significantly differ from the zero point. As would be expected due to the symmetry of the player roles, the actual outcomes for a "prosocial" policy interacting with a "selfish" policy do not impact the overall utility. Overall, these results clearly demonstrate the important roles that context and counterfactual evaluation play in analyzing the cooperativeness of different behaviors in a predator-prey environment.

6 Conclusion

We motivated and specified a framework for measuring cooperative behavior that is contextual, customizable, and counterfactually contrastive. The cooperativeness measure is defined on a broad class of games and is agnostic to the mechanisms that drive cooperation, making it applicable to a variety of agent models. We then evaluated this measure on policies of games of increasing complexity, showing that the measure works in accordance with our intuitions and is capable of precluding examples of non-cooperative group behavior that contingently provide a group benefit.

One possible limitation of our approach is that by making cooperativeness the property of a stochastic policy rather than a learning algorithm in the RL case, we don't take into account the possibility of the context agent adapting to the other agent–this would be more challenging to define due to the dynamic nature of returns in multi-agent RL. Nonetheless, a promising direction for future work is to apply our measure to analyze policies generated by cooperation-oriented learning algorithms. For instance, our framework could be used to quantitatively examine how systems trained with concepts like altruistic regret, as explored by [24], translate their learning objectives into behavior that is cooperative by our introduced metric.

Our framework's application extends naturally to the domain of security games, where the intentions of agents are often uncertain and outcomes can be misleading. In many security scenarios, from network defense to infrastructure protection, an adversary may seek to behave in a manner that appears cooperative or benign on the surface to avoid detection before striking [2]. A purely outcome-based measure might fail to identify such a threat. Our counterfactually contrastive metric, however, provides a more robust analytical tool. By evaluating an agent's policy against its selfish best-response baseline, the metric can quantify subtle deviations from truly cooperative behavior. For instance, an agent consistently choosing actions that align perfectly with its selfish interests, while providing some incidental group benefit, could be flagged as non-cooperative and worthy of further scrutiny.

Furthermore, the customizable and contextual nature of our measure is particularly well-suited for analyzing collusion and coordinated attacks, a central problem in security [57]. As noted, one can adjust the context to measure cooperativeness within a specific subgroup of agents. In a security game, this allows for the quantitative identification of potential adversarial coalitions. A high cooperativeness score within a subgroup, especially when that subgroup's actions are detrimental to the wider system's welfare, can serve as a formal signal for collusive behavior. This approach could be used to analyze the resilience of multi-agent systems against such threats or to develop adaptive defense mechanisms that monitor for the emergence of anomalously cooperative clusters of agents [28].

Acknowledgments. This work was supported by the John Templeton Foundation (grant number 62220). We are grateful for helpful conversations with: other members of the Laboratory for Intelligent Probabilistic Systems; Tom Griffiths and members of the Princeton Computational Cognitive Science Lab; and the 2023 Cooperative AI Summer School.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Bibliography

[1] Agapiou, J.P., Vezhnevets, A.S., Duéñez-Guzmán, E.A., Matyas, J., Mao, Y., Sunehag, P., Köster, R., Madhushani, U., Kopparapu, K., Comanescu, R., et al.: Melting Pot 2.0. arXiv preprint arXiv:2211.13746 (2022)

[2] Alpcan, T., Başar, T.: Network Security: A Decision and Game-Theoretic Approach. Cambridge University Press (2010)

[3] Arrow, K.J., Sen, A., Suzumura, K.: Handbook of Social Choice and Welfare, vol. 2. Elsevier (2010)

[4] Axelrod, R., Hamilton, W.D.: The evolution of cooperation. Science **211**(4489), 1390–1396 (1981)

[5] Bradbury, J., Frostig, R., Hawkins, P., Johnson, M.J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., Zhang, Q.: JAX: composable transformations of Python+NumPy programs (2018), `http://github.com/google/jax`

[6] Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., Graepel, T.: Cooperative AI: machines must learn to find common ground (2021)

[7] Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K.R., Leibo, J.Z., Larson, K., Graepel, T.: Open problems in cooperative AI. arXiv preprint arXiv:2012.08630 (2020)

[8] Dawes, R.M.: Social dilemmas. Annual Review of Psychology (1980)

[9] Du, Y., Leibo, J.Z., Islam, U., Willis, R., Sunehag, P.: A review of cooperation in multi-agent learning. arXiv preprint arXiv:2312.05162 (2023)

[10] Duéñez-Guzmán, E.A., Sadedin, S., Wang, J.X., McKee, K.R., Leibo, J.Z.: A social path to human-like artificial intelligence. Nature Machine Intelligence **5**(11), 1181–1188 (2023)

[11] Foerster, J.N.: Deep Multi-Agent Reinforcement Learning. Ph.D. thesis, University of Oxford (2018)

[12] Foerster, J.N., Farquhar, G., Afouras, T., Nardelli, N., Whiteson, S.: Counterfactual multi-agent policy gradients. In: McIlraith, S.A., Weinberger, K.Q. (eds.) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018. pp. 2974–2982. AAAI Press (2018). `https://doi.org/10.1609/AAAI.V32I1.11794`, `https://doi.org/10.1609/aaai.v32i1.11794`

[13] Hamilton, W.D., Hamilton, W.D.: Narrow roads of gene land: evolution of social behaviour, vol. 1. Oxford University Press on Demand (1996)

[14] Hertz, U., Koster, R., Janssen, M., Leibo, J.Z.: Beyond the matrix: Experimental approaches to studying social-ecological systems (2023)

[15] Hughes, E., Leibo, J.Z., Phillips, M., Tuyls, K., Dueñez-Guzman, E., García Castañeda, A., Dunning, I., Zhu, T., McKee, K., Koster, R., et al.: Inequity aversion improves cooperation in intertemporal social dilemmas. Advances in Neural Information Processing Systems **31** (2018)

[16] Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P., Strouse, D., Leibo, J.Z., De Freitas, N.: Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In: International Conference on Machine Learning. pp. 3040–3049. PMLR (2019)

[17] Kleiman-Weiner, M., Ho, M.K., Austerweil, J.L., Littman, M.L., Tenenbaum, J.B.: Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction. In: CogSci (2016)
[18] Kollock, P.: Social dilemmas: The anatomy of cooperation. Annual Review of Sociology **24**(1), 183–214 (1998)
[19] Koutsoupias, E., Papadimitriou, C.: Worst-case equilibria. Computer Science Review **3**(2), 65–69 (2009)
[20] Kropotkin, K.P.: Mutual aid: A factor of evolution. Black Rose Books Ltd. (2021)
[21] Leibo, J.Z., Hughes, E., Lanctot, M., Graepel, T.: Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research. arXiv preprint arXiv:1903.00742 (2019)
[22] Leibo, J.Z., Zambaldi, V., Lanctot, M., Marecki, J., Graepel, T.: Multi-agent reinforcement learning in sequential social dilemmas. In: Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems. pp. 464–473 (2017)
[23] Lerer, A., Peysakhovich, A.: Maintaining cooperation in complex social dilemmas using deep reinforcement learning. arXiv preprint arXiv:1707.01068 (2017)
[24] Loftin, R., Bandyopadhyay, S., Çelikok, M.M.: On the complexity of learning to cooperate with populations of socially rational agents (2024), `https://arxiv.org/abs/2407.00419`
[25] Luce, R.D., Raiffa, H.: Games and decisions: Introduction and critical survey. Courier Corporation (1989)
[26] Macy, M.W., Flache, A.: Learning dynamics in social dilemmas. Proceedings of the National Academy of Sciences **99**(suppl_3), 7229–7236 (2002)
[27] Mao, Y., Reinecke, M.G., Kunesch, M., Duéñez-Guzmán, E.A., Comanescu, R., Haas, J., Leibo, J.Z.: Doing the right thing for the right reason: Evaluating artificial moral cognition by probing cost insensitivity. arXiv preprint arXiv:2305.18269 (2023)
[28] Mazrooei, P., Archibald, C., Bowling, M.: Automating collusion detection in sequential games. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 27, pp. 675–682 (2013)
[29] Mordatch, I., Abbeel, P.: Emergence of grounded compositional language in multi-agent populations. arXiv preprint arXiv:1703.04908 (2017)
[30] Nowak, M., Sigmund, K.: A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. Nature **364**(6432), 56–58 (1993)
[31] Paternotte, C.: Minimal cooperation. Philosophy of the Social Sciences **44**(1), 45–73 (2014)
[32] Peña, J., Nöldeke, G.: Cooperative dilemmas with binary actions and multiple players (2023)
[33] Peng, B., Rashid, T., Schroeder de Witt, C., Kamienny, P.A., Torr, P., Böhmer, W., Whiteson, S.: Facmac: Factored multi-agent centralised policy gradients. Advances in Neural Information Processing Systems **34**, 12208–12221 (2021)
[34] Perolat, J., Leibo, J.Z., Zambaldi, V., Beattie, C., Tuyls, K., Graepel, T.: A multi-agent reinforcement learning model of common-pool resource appropriation. Advances in Neural Information Processing Systems **30** (2017)
[35] Rapoport, A., Chammah, A.M., Orwant, C.J.: Prisoner's dilemma: A study in conflict and cooperation, vol. 165. University of Michigan press (1965)
[36] Rutherford, A., Ellis, B., Gallici, M., Cook, J., Lupu, A., Ingvarsson, G., Willi, T., Khan, A., Schroeder de Witt, C., Souly, A., et al.: Jaxmarl: Multi-agent rl environments and algorithms in jax. In: Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems. pp. 2444–2446 (2024)

[37] Schelling, T.C.: The Strategy of Conflict: with a new Preface by the Author. Harvard University Press (1980)
[38] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
[39] Shapley, L.S.: Notes on the n-person game – ii: The value of an n-person game. Tech. Rep. RM-670-PR, RAND Corporation, Santa Monica, Calif. (August 1951)
[40] Shapley, L.S.: Stochastic games. Proceedings of the National Academy of Sciences **39**(10), 1095–1100 (1953)
[41] Shoham, Y., Leyton-Brown, K.: Multiagent systems: Algorithmic, game-theoretic, and logical foundations. Cambridge University Press (2008)
[42] Singer-Clark, T.: Morality metrics on iterated prisoners dilemma players (2014)
[43] Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction. MIT press (2018)
[44] Tan, M.: Multi-agent reinforcement learning: Independent vs. cooperative agents. In: Proceedings of the Tenth International Conference on Machine Learning. pp. 330–337 (1993)
[45] Tang, N., Gong, S., Zhao, M., Gu, C., Zhou, J., Shen, M., Gao, T.: Exploring an imagined "we" in human collective hunting: Joint commitment within shared intentionality. In: Proceedings of the Annual Meeting of the Cognitive Science Society. vol. 44 (2022)
[46] Tang, N., Stacy, S., Zhao, M., Marquez, G., Gao, T.: Bootstrapping an imagined we for cooperation. In: CogSci (2020)
[47] Tomasello, M.: Why we cooperate. MIT press (2009)
[48] Trivers, R.L.: The evolution of reciprocal altruism. The Quarterly Review of Biology **46**(1), 35–57 (1971)
[49] Tuomela, R.: What is cooperation? Erkenntnis pp. 87–101 (1993)
[50] Vinitsky, E., Köster, R., Agapiou, J.P., Duéñez-Guzmán, E.A., Vezhnevets, A.S., Leibo, J.Z.: A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings. Collective Intelligence **2**(2) (2023)
[51] Wang, J., Zhang, Y., Gu, Y., Kim, T.K.: Shaq: Incorporating shapley value theory into multi-agent q-learning. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 5941–5954. Curran Associates, Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/27985d21f0b751b933d675930aa25022-Paper-Conference.pdf
[52] West, S.A., Griffin, A.S., Gardner, A.: Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. Journal of Evolutionary Biology **20**(2), 415–432 (2007)
[53] Willis, R., Du, Y., Leibo, J.Z., Luck, M.: Resolving social dilemmas with minimal reward transfer. arXiv preprint arXiv:2310.12928 (2023)
[54] Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., Wu, Y.: The surprising effectiveness of PPO in cooperative multi-agent games. Advances in Neural Information Processing Systems **35**, 24611–24624 (2022)
[55] Yuan, L., Zhang, Z., Li, L., Guan, C., Yu, Y.: A survey of progress on cooperative multi-agent reinforcement learning in open environment (2023)
[56] Zhao, M., Tang, N., Dahmani, A.L., Zhu, Y., Rossano, F., Gao, T.: Sharing rewards undermines coordinated hunting. Journal of Computational Biology (2022)
[57] Zhou, C.V., Leckie, C., Karunasekera, S.: A survey of coordinated attacks and collaborative intrusion detection. Computers Security **29**(1), 124–140 (2010)

A Three-player Tabular Cleanup

The setup of the three-player version of Tabular Cleanup is the same, with the following changes to the policy to reflect the greater number of players:

- *Always* X: In this case, *Always Punish* punishes another player at random.
- *TFT*: In this case, *TFT* n will clean if n players are also cleaning, and will eat otherwise. Hence, TFT 2 is a more "suspicious" reciprocator than TFT 1 [1].
- *Nash*: This policy does not change, though it is worth noting that it is still a Nash equilibrium to the game if all players choose this policy.
- *Prosocial*: This policy also does not change, though it is no longer maximally prosocial in the three-player version. In fact, there is no maximally prosocial policy for this game that is symmetric across all player indices.

B Full results across multiple welfare functions for all tabular games

We evaluate the cooperativeness measure on each game using three different welfare metrics. Let $V_1, V_2, ..., V_N$ denote the values for N agents in the environment. These metrics are then defined as:

- Total Value: $\sum_{i=1}^{N} V_i$.
- Minimum Value: $\min_{i=1,2,...,N} V_i$.
- Equality: $1 - \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} |V_i - V_j|}{2N \sum_{i=1}^{N} V_i}$.

C Experiment Details

The experiments in Sec. 5 were run on Tyan Thunder servers, with an NVidia RTX A5000 GPU used for Sec. 5.4. Each training run (across all random seeds) takes no more than 15 minutes.

MARL models were trained with an MLP actor-critic architecture using the PPO algorithm [38], with the hyperparameters shown in Table 1.

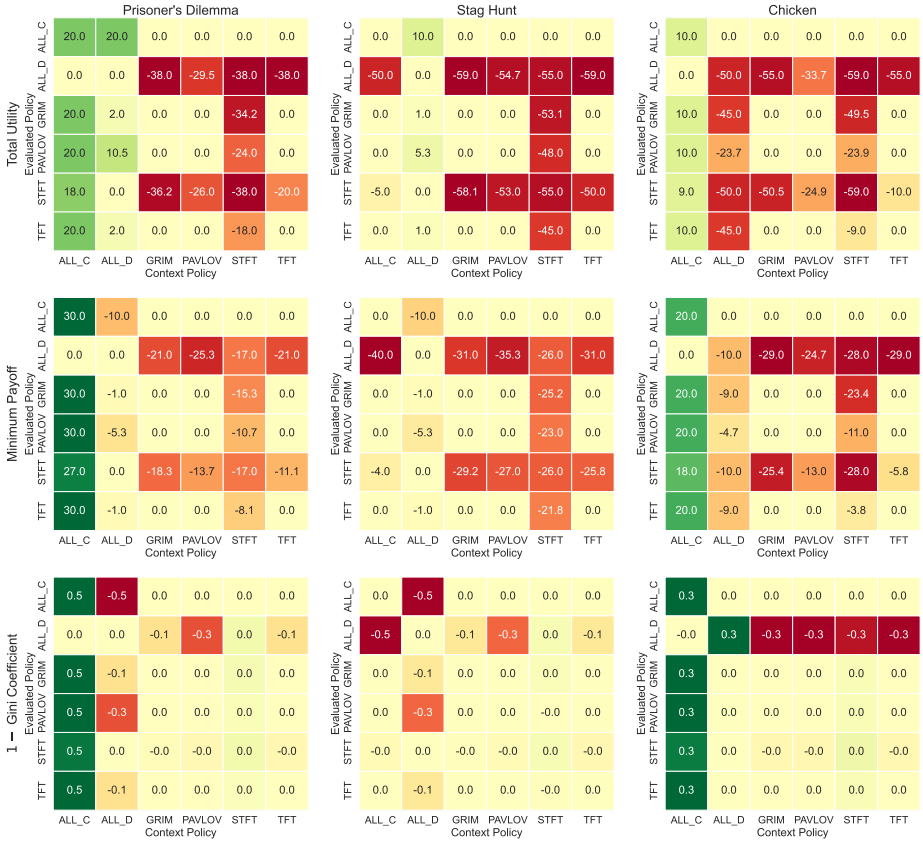

Fig. 7. Cooperativeness of six common deterministic policies in the Iterated Prisoner's Dilemma, in the context of each other policy. The cooperativeness is valued on the initial state, with a discount factor $\gamma = 0.9$.

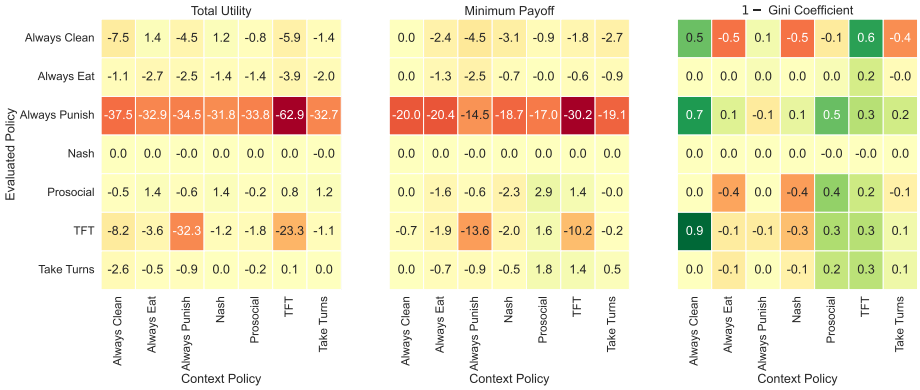

Fig. 8. Cooperativeness of seven deterministic policies in the 2-player Tabular Cleanup, in the context of each other policy, for all three welfare functions. The cooperativeness is valued on the initial state, with a discount factor $\gamma = 0.9$.

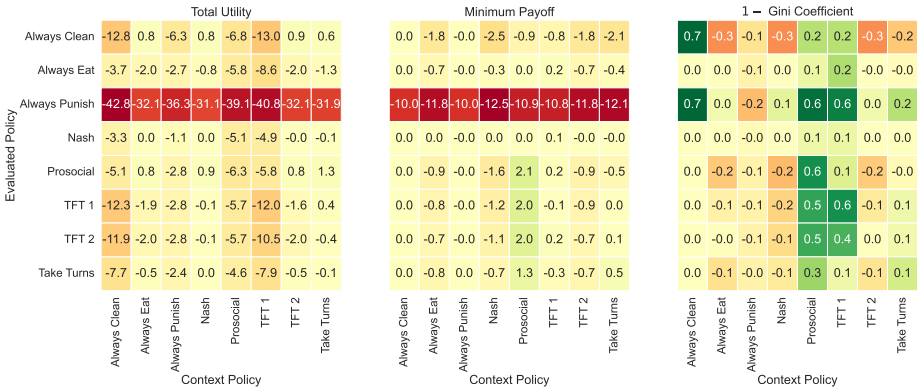

Fig. 9. Cooperativeness of eight deterministic policies in the 3-player Tabular Cleanup, in the context of two players playing the same context policy, for all three welfare functions. The cooperativeness is valued on the initial state, with a discount factor $\gamma = 0.9$.

Table 1. MARL training hyperparameters

Name	Value
Number of Hidden Layers	2
Layer Width	64
Layer Activation	tanh
Learning Rate (LR)	2.5e-4
Number of Steps	128
Total Timesteps	2e7
Update Epochs	4
Number of Minibatches	4
Discount Factor	0.99
GAE Lambda	0.95
Clip Epsilon	0.2
Entropy Coefficient	0.01
Value Function Coefficient	0.5
Max Gradient Norm	0.5
Anneal Learning Rate	True
Initial Random Seed	30
Number of Environments	16
Number of Seeds within each Environment	10