Learning to Penalize Other Learning Agents

Kyrill Schmid[1], Lenz Belzner[2] and Claudia Linnhoff-Popien[1]

[1]Ludwig-Maximilians-Universität München, Munich
[2]MaibornWolff GmbH, Munich
kyrill.schmid@ifi.lmu.de

Abstract

A key challenge in AI is the development of algorithms that are capable of cooperative behavior in interactions involving multiple independent machines or individuals. Of particular interest are social dilemmas, which are situations that raise tension between an individual's best choice and the desirable outcome in terms of the group. Although such scenarios have been studied increasingly within the AI community recently, there are still many open questions on which aspects drive cooperative behavior in a particular situation. Based on the insights from behavioral experiments that have suggested positive effects of penalty mechanisms towards cooperation, in this work we adopt the notion of penalties by enabling independent and adaptive agents to penalize others. To that end, we extend agents' action spaces with penalty actions and define a negative real-valued punishment value. We utilize reinforcement learning to simulate a process of repeated interaction between independent agents, learning by means of trial-and-error. Our evaluation considers different two player social dilemmas, and the N-player Prisoner's Dilemma with up to 128 independent agents, where we demonstrate that the proposed mechanism combined with decentralized learning significantly increases cooperation within all experiments.

Introduction

The field of Cooperative AI aims at identifying vital aspects that drive the emergence of cooperation in the interaction of multiple (independent) decision makers (Dafoe et al., 2020). The benefits of using AI to pursue this question are twofold: on the one hand finding methods that help to make AI systems capable of cooperative behavior is a mandatory step to broadly integrate and apply AI agents in our daily lives. On the other hand, by using AI to study the emergence of cooperation, one can approach outstanding questions from related fields for instance from game theory, psychology or economics. AI provides a rich set of tools to analyze complex models featuring many agents or spatial and temporal extended domains where it is hard to apply theoretical solution methods. Moreover, with AI agents one can design experiments that are not susceptible to human biases in decision making, but are driven by objectives such as individual reward maximization, which is hard to study in experiments featuring humans.

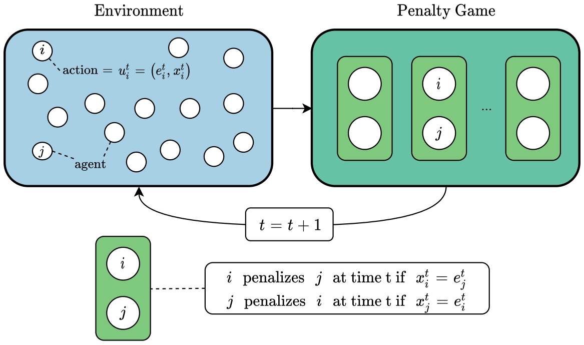

Environment
Penalty Game
i
action $= u_i^t = (e_i^t, x_i^t)$
j
agent
i
j
...
$t = t + 1$
i
j
i penalizes j at time t if $x_i^t = e_j^t$
j penalizes i at time t if $x_j^t = e_i^t$

Figure 1: Idea of the penalty mechanism: at each step of the iterated game, agents are matched randomly. Pairs of agents then have the possibility to penalize their respective actions. If agent i penalizes action e_j^t from agent j, and j actually executed e_j^t at t, then the punishment value p will be subtracted from j's reward and added to i's.

The question what causes cooperation between multiple players has been studied extensively in the field of game theory on the basis of so called social dilemmas, which allow to study the emergence of cooperation in abstract models represented as matrix games. Each social dilemma poses a challenge for the emergence of cooperation, as there is tension between the best choice in terms of the collective, i.e. what is best for all players, and the choice that is individually rational (Ostrom, 1990, 2008). Social dilemmas can be categorized into models featuring two agents and N-player models that are not limited in terms of players. With two players and two actions, the payoffs can be represented in a 2×2 matrix, where cell i, j contains the payoffs X and Y for both players as a consequence of action i, j (see Figure 2. Most famously, in the Prisoner's Dilemma each agent has an incentive to defect, either to exploit the (cooperative) other player or because of fearing of being exploited (for being cooperative) by the other player. Therefore, the only (non Pareto-optimal) equilibrium is given by mutual defection, which renders the dilemma. With more than two players, the scenarios can be further categorized in *public good*

Stag Hunt	C	D
C	4, 4	0, 3
D	3, 0	1, 1

Chicken	C	D
C	3, 3	1, 4
D	4, 1	0, 0

Prisoners	C	D
C	1, 1	-0.5, 1.5
D	1.5, -0.5	0, 0

Figure 2: Three canonical two player social dilemmas: *Chicken*, *Stag Hunt*, and the *Prisoner's Dilemma*.

games and *common pool* games. The former describes situations where agents need to jointly invest into a public good in order to enjoy the benefits of the public good. Examples for public good games can be found in biology, such as predator inspection behavior or group defense, as well as in human societies, e.g. health insurance or public transportation. In public good scenarios, individuals are incentivized to freeride, i.e. not contribute to the provision of the public good but still enjoy the benefits of the public good. Eventually, freeriding can lead to the non-provisioning of the public good if to few individuals contribute. However, there is no rivalry for the public good between players. In contrast in common pool games, there is a resource that can be used by all individuals but there is rivalry between agents for the resource, as the overall utility decreases through individual usage. Here the tragedy arises due to overusage, as the costs of using the resource are beard by all individuals, whereas the benefits are earned on an individual basis.

In this work, we study the problem of multiple independent learning agents in different well known types of two player and N-Player social dilemmas. Whereas earlier work in the field of multi-agent reinforcement learning studied the influence on cooperation of various parameters such abundance in resources (Leibo et al., 2017) or spatial features of the domain (Perolat et al., 2017), here we focus on an aspect associated with agents' capabilities to regulate each other. More specifically, in this work we use a penalizing mechanism that enables individual agents to punish others. The idea of letting players penalize each other has been addressed earlier, both theoretically and in experimental research, which found a positive impact of penalties towards the general willingness to cooperate (Janssen et al., 2010; Ostrom et al., 1992). However, prior work defined penalties as an operation that produces costs for those who penalize others. This circumstance can pose a second order social dilemma (Kollock, 1998): if the act of penalizing other agents is associated with costs, be it through the time that is spent to impose a punishment which can not be spend to increase one's own utility (Perolat et al., 2017), or the actual payoff that needs to be invested to penalize someone (Ostrom et al., 1992), then players might decide to leave the expense of punishing to others, i.e. decide to freeride at the cost of others.

Here we propose a method to solve both dilemmas simultaneously:

- By enabling players to penalize other players, defective behavior will become less tempting so players become more cooperative.
- Agents who successfully impose a penalty can achieve a personal benefit by earning a payoff that is subtracted from from the penalized player, so agents are incentivized to make use of the penalty mechanism.

We find in different experiments that such an integrated mechanism stably promotes cooperative policies for multi-agent systems involving more than 100 learning agents. We also demonstrate that the usage of the penalty actions defines a dominant strategy in the game, as an agent which learns to penalize can achieve a higher reward than an agent who never makes use of the penalty mechanism. The code for the experiments in this paper can be found on github [1].

Fundamentals

A N-player normal form game, denoted Γ, is defined as a three tuple $(N, (\mathcal{A}_i)_{i\in N}, (r_i)_{i\in N})$, where N is the number of agents, $\mathcal{A}_i$ is player i's set of actions, and $r_i : \Pi_{i\in N}\mathcal{A}_i \rightarrow \mathbb{R}$ is player i's payoff (also called reward) function. For the scope of this work, we consider games with two actions for all players, referred to as cooperation C and defection D. Depending on the specific rewards for agents, a game can be categorized into three classes (Dafoe et al., 2020): 1) *pure common interest games* where an increase in an agent's payoff also increases the payoff of all others, 2) *mixed-motive games* are scenarios with general sum rewards, so agents might either have conflicting or aligned goals, 3) *pure conflicting goals* in which an increase in one agents payoff is always associated with a decrease in the reward of others. Games of pure common interest and mixed-motive games can principally have opportunities for agents to cooperate, thereby increase both, the overall reward and their individual rewards.

2-Player Social Dilemmas Interactions between two players are an important class of games to analyze vital aspects of cooperation by means of game theoretic solution methods like the Nash equilibrium. Three of the most popular canonical examples for 2 player mixed-motive games, are

[1] `https://github.com/kyrillschmid/penalty-games`

the Prisoner's Dilemma (PD), the Chicken game (CH), and Stag Hunt (SH), with rewards for agents displayed in Figure 2. Each of the three games raises conflict between players, motivated either by fear or greed or both: In Chicken, both players can profit from mutual cooperation, yet each player can be better off by unilaterally defecting, so a player might decide to defect out of greed. In Stag Hunt, the players can get a high reward by mutual cooperation. However, in case of only unilateral cooperation, the cooperative player receives nothing, whereas the defective player will get a positive reward, so a risk-averse player might decide to defect by default. Finally, in the Prisoner's Dilemma, players have an incentive to defect out of fear or greed, as they can personally benefit from unilateral defection while at the same time are in danger to become exploited by the other player for being cooperative. Note that two player games can be used in order to analyze interaction with more than two players by playing the game iteratively and at each iteration of the game different agents are matched with one another to play the 2-player matrix game. The matching of players can be done on the basis of different schemes, e.g. randomly or in a round robin fashion as proposed in the Prisoner's Dilemma tournament (Axelrod and Hamilton, 1981).

N-Player Social Dilemmas Games involving more than two agents can be categorized by the way costs and benefits are distributed between the players (Kollock, 1998). The first category, called public good games, is defined by games where the players need to incur costs in order to realize a public good, that is non-rival and from which players cannot be excluded. All players receive a benefit (positive reward) if the public good is provided but individuals are inclined to avoid their own costs by not participating. It is therefore that these scenarios are apt to produce pathological freeriding problems, which might even lead to the non-provisioning of the public good when too few players decide to participate. An instance of a public good game is described by the N-player Prisoner's Dilemma, which has the following characteristics: 1) each player has a dual choice (cooperate C or defect D), 2) the positive reward for defection goes to the defective player, while the cost of defection is distributed among all, 3) the overall reward increases with the number of cooperators, 4) the defective strategy is dominant (Edney and Harper, 1978).

In the second category, called common pool games, agents can get an immediate benefit from using a non-excludable but rival resource, so agents impose costs to others by using the resource. In these scenarios agents have incentive to overuse the shared resource, as the marginal benefit from using another unit are earned individually, while the marginal costs of using another unit are shared between all players. Common pool games give rise to the tragedy of the commons (Ostrom, 2008), that describes the situation where multiple individuals have access to a common and depletable resource such as pastures or fishing grounds. The tragic lies in the gap between individual and collective rationality: collectively, it is desirable to use the resource only up to a certain degree that allows it to regenerate over time. Individually however, it is rational to overuse the resource as the benefits of another taken unit are earned individually, but the costs for this unit are carried by the whole group.

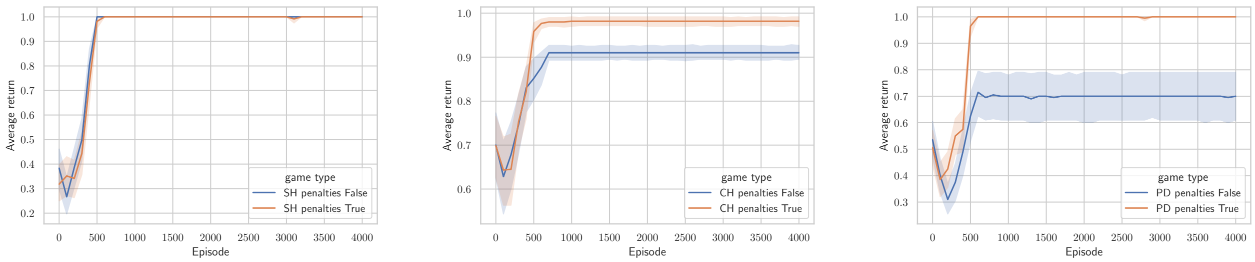
There are a number of differences regarding the learning dynamics in games between two players and games with more than two players (Dawes, 1980; Kollock, 1998). First, in two player games, agents might be able to infer the other player's action through their own rewards. This is not necessarily the case with more than two agents, where players might defect rather unnoticed. Second, in case of two players agents impose costs (or benefits) directly to each other through their choice, whereas costs are distributed among a potentially large collective with N players. The last point concerns the influence agents have upon each other: in a two player game, a player can shape the payoffs from its opponent by strategically choosing its own actions. It thereby can significantly shape the opponents behavior through its own behavior.

Method

In this section we describe our proposed penalization mechanism. Our approach is inspired by evidence which suggests that humans display the ability to overcome the bad outcomes predicted from non-cooperative game theory (Ostrom, 1990), which can be also be explained in theory through models that include sanctioning mechanisms (Ostrom et al., 1992). Key factors for emergent cooperation between humans are commitments, mutual monitoring of behaviors, and the possibility to impose penalties on those who display defective behavior (Ostrom, 1990). Here, we build on these insights by enabling agents to make use of a penalization mechanism so as to discourage defective behavior, which in turn means to increase cooperation. For this approach we extend a given N-player normal form game with a set of penalty actions, and a penalty value and call this extended game a *Penalty game* (see Figure 1), which we formally define as a tuple $(\Gamma, (\mathcal{A}_i^s)_{i \in N}, p)$, where

- Γ is the underlying normal form game, consisting of N agents, $\mathcal{A}_i$ is player i's set of actions, and the reward functions $r_i : \Pi_{i \in N} \mathcal{A}_i \rightarrow \mathbb{R}$ for each agent $i \in N$.
- The set of penalty actions $\mathcal{A}_i^p$ for each agent $i \in N$.
- The penalty value $p \in \mathbb{R}_{<0}$

In the Penalty game, the action space is extended for each agent with a fixed number of penalty actions $\mathcal{A}^p$, that provide the tool to penalize other agents with the penalty value p. Agents in the Penalty game, choose a so called environmental action e from the original action space $\mathcal{A}$, and



(a) Stag Hunt (b) Chicken (c) Prisoner's Dilemma

Figure 3: Results for the two player social dilemmas for agents with the penalization mechanism (orange) and agents without penalties (blue). (Mean and 95% confidence interval, best viewed in color)

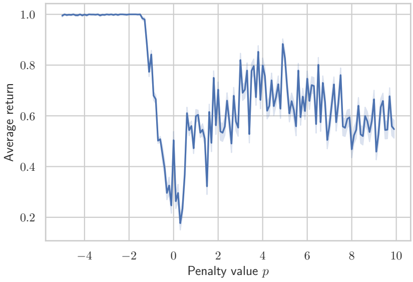


Figure 4: Normalized reward in the iterated Prisoner's Dilemma for different penalty values p averaged over 25 independent runs for each value.

a penalty action x from the penalty action set $\mathcal{A}^p$. The combined action space therefore has size $|\mathcal{A} \times \mathcal{A}^p|$. The fine granularity with which agents can specify whom they target and which action they intend to penalize, can be controlled via the semantic of the penalty actions. Without any restrictions, agent i can penalize any other agent $j \in N$ for any specific action $u \in \mathcal{A}$. In this case, agent i's action is a tuple $u_i = (e_i, x_{i,j})$, where $e_i \in \mathcal{A}$ is i's executed environmental action and $x_{i,j} \in \mathcal{A}^p$ defines the action to be penalized from agent j. The target agent j will be forced to pay the penalty p, if its environmental action e_j is equal to the defined penalty action $x_{i,j}$, in which case the rewards for i and j are:

$$r'_i(..., x_{i,j}, ..., e_j, ...) = r_i + |p * \delta_{x_{i,j}, e_j}|$$

$$r'_j(..., x_{i,j}, ..., e_j, ...) = r_j - |p * \delta_{x_{i,j}, e_j}|$$

where $\delta_{i,j}$ is the Kronecker delta with $\delta_{i,j} = 1$ if $i = j$ and 0 else.

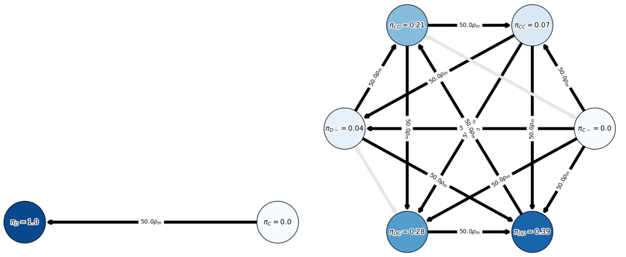
Note that when agents can penalize all other agents for any specific action the increase in action space is growing exponentially with the number of agents, since the size of the action space for agent i is then defined by $|\mathcal{A}_i| = |\mathcal{A}_i| * \Pi_{j \in N-1} |\mathcal{A}^p_j|$. In this work we therefore take an approach to effectively reduce the complexity for growing numbers of agents. To that end we propose to match two agents randomly at a time, such that these two agents have the chance to penalize each other. In this case the growth of the action space is constant in the number of agents since $|\mathcal{A}_i| = |\mathcal{A}_i \times \mathcal{A}^s_i|$. For the iterated version of the social dilemmas used for evaluation, this means that at each step agents are matched with a new partner at each step of the iterated game.

Learning To learn strategies we utilize reinforcement learning (RL), which refers to methods that learn in a trial-and-error based way. An agent's goal is to learn a policy π that maximizes its expected return $\mathcal{R}_t := \sum_{t=1}^{\infty} \gamma^{t-1} R_t$ where $\gamma < 1$ is a discount factor (Sutton and Barto, 2018). One way to learn a policy is to learn an action-value function $Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, where $Q(s, a)$ represents the value of action a in state s. The action-value function can be used as a policy by selecting actions according to their action values. A popular way to learn the action-value function is Q-learning, where an agent i updates its values according to:

$$Q_i(s, a) \leftarrow Q_i(s, a) + \alpha \big[r_i + \gamma \max_{a' \in \mathcal{A}^i} Q_i(s', a') - Q_i(s, a) \big]$$

where α is the learning rate and γ is a discount factor. During training, exploration can be incorporated through so called epsilon greedy action selection, where an agent selects its optimal action according to its current Q-function with probability $1 - \epsilon$ or a random action with probability ϵ. In this work, we represent the state in the iterated versions of social dilemmas either as a constant (0 in case of two agents) or the fraction of cooperators in the last step (for the N-player social dilemmas).

We model each player in the game as an independent instance of a tabular Q-learner, which is known as independent Q-learning. Although independent learning is known to render the learning problem non-stationary from a single agent's perspective, it is a natural way to model learning



(a) No penalties

(b) With penalties

Figure 5: Change in the α-Rank discrete-time dynamics of the Prisoner's Dilemma in the standard version and with penalties enabled: Without penalties the only stable strategy is to defect D (i.e. single Nash-equilibrium). With penalties enabled, pure defection becomes unattractive, while the fitness of strategies including penalties increases.

in case of mixed-motive games (Leibo et al., 2017; Perolat et al., 2017), since agents might have conflicting goals and therefore are unlikely to jointly optimize a centralized objective function. Moreover, the independent learning paradigm has been shown to outperform state-of-the-art centralized approaches such as COMA or QMIX (Mahajan et al., 2019).

For the training of N agents, each agent updates its Q-function based on its experienced transitions that comprise actions, rewards and next states, i.e. (s, a, r, s') to update its Q-function. We model episodic learning by considering iterated versions of the normal form games. In this work, an iterated game comprises 4000 consecutive steps (also called episodes), before the game restarts. We use a learning rate of $\alpha = 0.2$ for the two player social dilemmas, and a learning rate $\alpha = 0.008$ for the N-player game. We linearly anneal the exploration constant ϵ over the course of all steps, starting with $\epsilon_0 = 1.0$ and decreasing it until $\epsilon_{4000} = 0.0001$. In all experiments we use a discount of $\gamma = 0.9$.

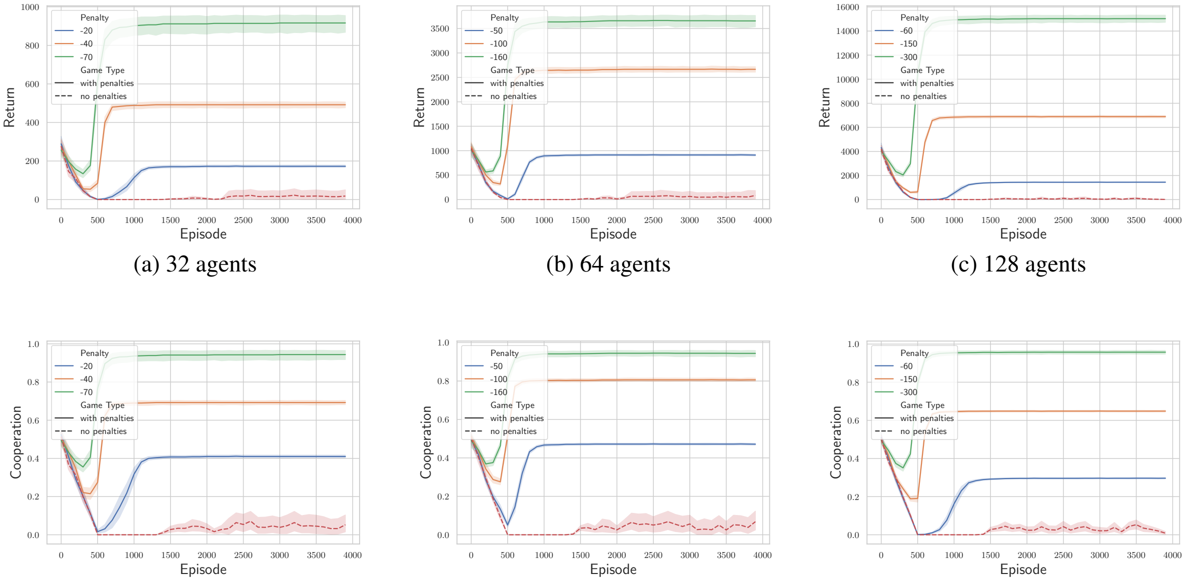
Results

In this section we describe experiments from two player social dilemmas, including Stag Hunt, Chicken and the Prisoner's Dilemma before extending the evaluation to a N-Player Social dilemma featuring up to 128 agents.

Two Players To study the effect of the proposed mechanism in two player social dilemmas we use three canonical matrix games known as the game of Chicken, the Prisoner's Dilemma, and Stag Hunt. It is known that by playing two player social dilemmas in an iterated fashion, cooperative strategies such as Tit-for-Tat which are based on reciprocity can thrive (Axelrod and Hamilton, 1981). However, in trial and error based learning such as reinforcement learning, cooperation based on reciprocity is unlikely to emerge since it requires recursive reasoning about the consequences of one's own behavior on others, which is not part of model-free RL. This circumstance can be seen in Figure 3, where all of the three games are played for 4000 consecutive steps, with results averaged over 100 independent runs (returns are normalized between 0 and 1). In Stag Hunt, independent Q-learning achieves near maximum overall reward, despite the existence of an non-optimal Nash-equilibrium. Chicken poses a harder challenge for cooperation due to its incentives for unilateral defection, which manifests in an overall decreased return. In the Prisoner's Dilemma independent Q-learning is likely to converge to the unique non-optimal Nash-Equilibrium (D, D) in some of the runs, which decreases overall return.

We now introduce the described penalty mechanism in the following way: In each of the three games, we extend agents' action spaces with additional action to let agent i penalize agent j based on j's played action, thus the action space is extended from $\{C, D\}$ to $\{(C, -), (C, C), (C, D), (D, -), (D, C), (D, D)\}$, where the first component indicates the player's own action and the second component is the intended punishments for the other player. Through this extension, as shown in Figure 3, the learned strategies by independent Q-learning can be improved towards near optimal play in Chicken and optimal play, that is full cooperation in the Prisoner's Dilemma. The difference in outcomes is strongest in the Prisoner's Dilemma, where strategies are consistently changed to mutual cooperation (C, C) after approximately 500 steps, thereby achieving the maximum reward. To define the optimal penalty value p, we considered values in the interval $[-5, 10]$. Overall we found, that a value of $p = -2.0$ achieved the best results in all three games. We also tested the effect of using positive values $p > 0$ (so agents are not punished but rewarded) and found that it rendered the learning dynamics more unstable but led to increased cooperation for some values while some positive values led to little or no cooperation (see Figure 4).

To illustrate the change in the dynamics of the game, we inspect the Prisoner's Dilemma by means of α-Rank, a population based evaluation technique (Omidshafiei et al., 2019). The α-Rank discrete time dynamics of the Prisoner's Dilemma with and without penalizing actions are shown in Figure 5. Without penalties, there are two actions (C, D) and the two graph nodes correspond to the situation where all individuals in the population play either cooperative or defective. The time the populations spend in each strategy is quantified as the mass of the stationary distribution in this node. Edges between nodes correspond to the fixation probabilities for state pairs, so edge directions indicate the flow of individuals from a strategy towards a fitter strategy. In the standard Prisoner's Dilemma all probability mass is accumulated in D, which means that defection is the fittest strategy, with no chance for cooperative individuals to survive. This is in compliance with the Nash-equilibrium for



(a) 32 agents (b) 64 agents (c) 128 agents

(d) 32 agents (e) 64 agents (f) 128 agents

Figure 6: Overall reward (sum of agent rewards) and cooperation ratio (share of cooperators at each step) for 32, 64 and 128 agents in the N-player public good game. Shown are mean and 95% confidence interval.

the Prisoner's Dilemma. The game dynamics change for the Prisoner's Dilemma with penalty actions, where more probability mass is assigned to the strategies DC, CD, CC, DD, i.e. strategies that include a penalty component. Moreover, there are circles which include cooperative strategies, so agents have more possibility to mutually adapt to a cooperative strategy. Although the circle indicates that there is no stable Nash-equilibrium, learning stably converges to the mutually efficient cooperative outcome, which might be due to the decreasing exploration rate of the Q-learning agents. With respect to the results of independent Q-learning in Figure 3, these changes of the game dynamics are sufficient to direct the learning process towards the globally optimal outcome.

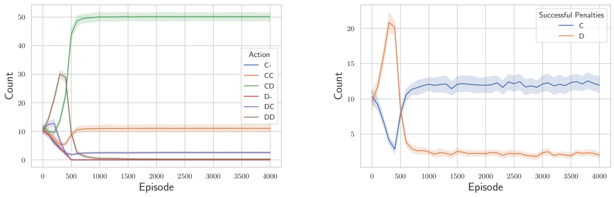
N Player Results We now extend the evaluation towards games involving more than two agents for which we consider $N \in \{32, 64, 128\}$. We utilize the N-player public good game proposed in (Barbosa et al., 2020) that has the following properties: Each cooperator contributes a positive amount P to the public good, whereas defectors do not contribute. The aggregated contribution from all cooperators is evenly distributed among all group members, but only cooperators bear the costs of providing the public good so defectors can benefit from the good at no cost. The reward functions for cooperators and defectors are:

$$R(C) = \frac{f * c * P}{N} - P, R(D) = \frac{f * c * P}{N}$$

where f is a constant, P is the amount of the provided public good, and c is the number of cooperators (note that we renamed the the parameter P to avoid confusion with the parameters defined in the Penalty game). For all runs we used the following parameters to compute the returns: The contribution p that each cooperative agent achieves is $P = 1$. For the scaling factor f we use $f = 2$ as described in (Barbosa et al., 2020).

First, we consider the overall reward (sum of all agent rewards) and the degree of cooperation that is achieved by agents with and without the penalty mechanism. Figure 6 shows results for the reward and the rate of cooperation rate within the population for each step ($\textit{number cooperators}/N$). For the evaluation we use different values for the penalty value p with $p \in [-20, 300]$. For all numbers of agents N, we find that the overall reward and the overall cooperation rate significantly increase with penalties being enabled. Moreover, the outcome depends on the amount of the penalty, where higher penalties (lower values for p) correspond to higher overall rewards and higher degrees of cooperation. When penalties are available, the overall cooperation ratio reaches levels above 90 percent for all settings, whereas without penalization stable cooperation fails as indicated by the low rewards and the ratio of cooperators.

We now want to consider which action in the game is actually played by agents and how many penalties were imposed successfully. The results are visualized in Figure 7, where



(a) Played action

(b) Successful imposed penalties

Figure 7: Played actions and number of successful penalties for 64 agents in the N-player public good game ($p = -160$). Shown are mean and 95% confidence interval.

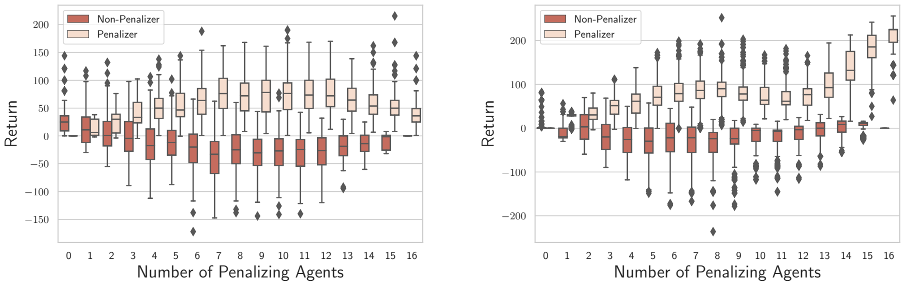
we consider 64 agents over the course of 4000 iterated steps in the game averaged over 25 independent runs. The number of actually played actions are visualized in Figure 7a. Plotted are all 6 actions of the Penalty game, where the label XY corresponds to action X that is executed in the game, and Y corresponds to the action to be penalized. So for instance CD means that the agent cooperates while it imposes a penalty for defecting. The results show that in the early stage of learning (i.e. for *episode* < 500), there is a strong increase in DD, which means that agents decide to defect while at the same time punish others for defecting. After this phase, action CD (being cooperative and punishing defection) becomes increasingly attractive and is consistently chosen for the rest of the training from around 50 out of 64 agents. There remains a number of agents which decide to cooperate and punish others for cooperating (around 10 agents) and a small group that chooses to defect while punishing others for cooperation (4 agents on average). This behavior is mirrored in the number of successful penalties, displayed in Figure 7b. In the beginning there is a sharp increase in successful penalties for defection D, which then decreases as there are only very little defectors left, so penalizing defectors becomes unattractive. Again, the small group of agents which is specialized in penalizing cooperators results in a constant share of successful penalties for cooperation after around episode 500.

Finally, we visualize the dynamics of the N-player public good game by means of empirical game theoretic analysis with a Shelling diagram. We therefore simulate runs with different numbers of agents who can penalize, denoted $|S|$, in a game comprising 16 agents in total. We then compare the aggregated rewards from all penalizing agents with the rewards from non-penalizing agents. Figure 8 shows the results. At the leftmost point in the plot there is no penalizing agent, so it resembles the original game. At the rightmost point, the game contains only penalizing agents, which corresponds to the game with penalties. We compare Shelling diagrams from an early training phase (episode < 500) with a Shelling diagram collected during late training (episode $>$ 3000) to see whether the dynamics of the game change for different training phases. During late training, the rewards of agents who can penalize others are higher for all numbers of penalizing agents in the game. This indicates that it is at any point individually rational to use the penalty mechanism if available. With more than half of all agents using penalties ($|S| > 16$), the rewards for all penalizing agents decreases slightly, until it recovers for $|S| > 11$. Consequently, the Penalty game displays a social dilemma (i.e. a common good game) for intermediate numbers of agents, since between 8 and 12 penalizing agents the return from all penalizing agents decreases when more penalizing agents enter the game. However, this effect seems locally limited, as with all agents being capable to penalize others the rewards of all agents increase and show the highest overall return.

Related Work

Positive effects of incentivation towards cooperative behavior in social dilemmas have been identified in the literature and can be distinguished in selective incentives and sanctioning mechanisms (Kollock, 1998). Selective incentives describe approaches that try to positively promote cooperation, e.g by giving monetary rewards to reduce the consumption of common pool goods, such as water or electricity (Maki et al., 1978; Winett et al., 1978). In contrast, incentivations that actively try to reduce defection work on the basis of penalties, for which experiments with humans suggest that penalties are effective in reducing defective behavior (Caldwell, 1976; Komorita, 1987). Whereas in experiments involving humans penalty systems have been realized by allowing participants to pay a fee in order to penalize other defective players (Janssen et al., 2010), work in the field of multi-agent reinforcement learning has adapted another approach to penalize other agents: in (Perolat et al., 2017), agents live in a grid-world and can increase their reward by gathering a shared resource (apples). The penalty mechanism in the gathering game works on the basis of a beam action, which bans the agent that is caught in the beam for 25 consecutive steps from the game, during which they cannot collect apples to increase their reward. In our work, agents influence other agents' rewards rather than penalizing others indirectly through a imposed time penalty, so the penalties trigger a transaction of reward between the innvolved agents.

Other work in the field of reinforcement learning, analyzed which environmental or agent internal parameters drive the emergence of cooperation between two players. For that, (Leibo et al., 2017) extend the notion of social dilemmas with so called sequential social dilemmas (SSDs) to better capture aspects of real world social dilemmas, as real world dilemmas are in general temporally extended and may feature non-binary grades of cooperation. The authors demonstrate that cooperation depends on different aspects of the environment such as the abundance of the shared re-



(a) Early training

(b) Late training

Figure 8: Shelling diagrams for the public good game with 16 RL agents. Displayed are the rewards for different numbers of agents who can penalize others. Left side: early training (episode < 500), right side: late training (episode > 3000).

source between agents. While the analysis demonstrated that specific variables influence cooperation, there is no specific proposal on how to increase cooperation for a given scenario. More recently, in (Yang et al., 2020) an approach has been proposed where agents learn an additional incentivation function with which they can increase the return of the other agent. This mechanism is different to our approach since the incentivation function does not define an economic transaction as the incentive an agent receives is not subtracted from the other agent's reward. Here, we leave the overall reward constant, as each transaction marks a bilateral exchange that equals in total. Other work, in the line of SSDs aims at establishing cooperation by incorporating social preferences such as inequity aversion into the model (Hughes et al., 2018), where it is shown that inequity-averse agents improve the temporal credit assignment problem and promote cooperation. An example that works with positive incentivations is given in (Lupu and Precup, 2020), an approach that allows agents to directly assign reward to other agents to overcome the tragedy of the commons. In (Schmid et al., 2018), so called action markets are proposed, where agents can learn to incentivize others through positively rewarding each other. This approach relates to the mechanism in this work, as rewards can be given conditionally on specific actions of agents. It differs through the positive incentivation value, which prohibits its application in social dilemmas, where the overall goal is mutual cooperation.

Opponent modelling (He et al., 2016; Raileanu et al., 2018; Everett and Roberts, 2018) has also been utilized in order to establish cooperation in SSDs, such as in (Wang et al., 2018), where a cooperation degree detection network was trained to identify the opponent's current level of cooperation. Based on the opponent's behavior, an agent can then select its response. Other work with the aim of building an opponent model allows agents to reason over the learning of other agents by an additional term within the learning rule (Foerster et al., 2018). The authors demonstrate that the encounter of two such agents can lead to tit-for-tat, a strategy famous for its cooperativeness and robustness regarding exploitation from defectors. In this work, agents do not build an opponent model, nor does learning involve any kind of recursive or theory of mind (Rabinowitz et al., 2018) like reasoning. Rather, other agents are considered as part of the environment by individual agents, such that results can be understood as the emergent outcome of independently learning agents trying to adapt to an ever changing environment. Finally, agents in this work do not communicate explicitly such as in (Foerster et al., 2016). Here, information is transferred only indirectly via agents' penalizing activity.

Conclusion

In this work we consider the problem of multi-agent learning in environments where agents can either be cooperative to increase the overall return or be defective to increase their individual payoff. These scenarios include well known two player social dilemmas like the Prisoner's Dilemma, Stag Hunt or Chicken but applies also to games involving potentially large numbers of agents, such as the N-player public good game. Inspired by theoretical findings and behavioral experiments, which assign positive effects from sanctioning mechanisms towards cooperation, we propose a penalty mechanism to tackle two challenges: solving the first order social dilemma through the direct effect of penalties, and incentivizing agents to become effective punishers to prevent the potential second order social dilemma. For the evaluation we model agents as independent Q-learners, interacting pairwise in an iterated version of the respective social dilemma. From experiments in two agent social dilemmas we find that the proposed penalties can achieve full cooperation in dilemmas where Q-learning without penalties fails to achieve cooperation and where game theory predicts mutual defection (Prisoner's Dilemma). Moreover, in N-player scenarios we find that penalizing agents achieve more than 90% cooperation in games with up to 128 agents compared to small and unstable rates of cooperation achieved by Q-learning without penalties.

References

Axelrod, R. and Hamilton, W. D. (1981). The evolution of cooperation. *science*, 211(4489):1390–1396.

Barbosa, J. V., Costa, A. H. R., Melo, F. S., Sichman, J. S., and Santos, F. C. (2020). Emergence of cooperation in n-person dilemmas through actor-critic reinforcement learning.

Caldwell, M. D. (1976). Communication and sex effects in a five-person prisoner's dilemma game. *Journal of Personality and Social Psychology*, 33(3):273.

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

Dawes, R. M. (1980). Social dilemmas. *Annual review of psychology*, 31(1):169–193.

Edney, J. J. and Harper, C. S. (1978). The commons dilemma. *Environmental Management*, 2(6):491–507.

Everett, R. and Roberts, S. (2018). Learning against non-stationary agents with opponent modelling and deep reinforcement learning. In *2018 AAAI spring symposium series*.

Foerster, J., Assael, I. A., de Freitas, N., and Whiteson, S. (2016). Learning to Communicate with Deep Multi-Agent Reinforcement Learning. In *Advances in Neural Information Processing Systems*, pages 2137–2145.

Foerster, J., Chen, R. Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., and Mordatch, I. (2018). Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 122–130. International Foundation for Autonomous Agents and Multiagent Systems.

He, H., Boyd-Graber, J., Kwok, K., and Daumé III, H. (2016). Opponent Modeling in Deep Reinforcement Learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1804–1813.

Hughes, E., Leibo, J. Z., Phillips, M., Tuyls, K., Dueñez-Guzman, E., Castañeda, A. G., Dunning, I., Zhu, T., McKee, K., Koster, R., et al. (2018). Inequity aversion improves cooperation in intertemporal social dilemmas. In *Advances in neural information processing systems*, pages 3326–3336.

Janssen, M. A., Holahan, R., Lee, A., and Ostrom, E. (2010). Lab experiments for the study of social-ecological systems. *Science*, 328(5978):613–617.

Kollock, P. (1998). Social dilemmas: The anatomy of cooperation. *Annual review of sociology*, 24(1):183–214.

Komorita, S. S. (1987). Cooperative choice in decomposed social dilemmas. *Personality and Social Psychology Bulletin*, 13(1):53–63.

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., and Graepel, T. (2017). Multi-Agent Reinforcement Learning in Sequential Social Dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 464–473. IFAAMAS.

Lupu, A. and Precup, D. (2020). Gifting in multi-agent reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 789–797.

Mahajan, A., Rashid, T., Samvelyan, M., and Whiteson, S. (2019). Maven: Multi-agent variational exploration. *arXiv preprint arXiv:1910.07483*.

Maki, J. E., Hoffman, D. M., and Berk, R. A. (1978). A time series analysis of the impact of a water conservation campaign. *Evaluation Quarterly*, 2(1):107–118.

Omidshafiei, S., Papadimitriou, C., Piliouras, G., Tuyls, K., Rowland, M., Lespiau, J.-B., Czarnecki, W. M., Lanctot, M., Perolat, J., and Munos, R. (2019). α-rank: Multi-agent evaluation by evolution. *Scientific reports*, 9(1):1–29.

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge university press.

Ostrom, E. (2008). Tragedy of the commons. *The new palgrave dictionary of economics*, 2.

Ostrom, E., Walker, J., and Gardner, R. (1992). Covenants with and without a sword: Self-governance is possible. *The American Political Science Review*, pages 404–417.

Perolat, J., Leibo, J. Z., Zambaldi, V., Beattie, C., Tuyls, K., and Graepel, T. (2017). A multi-agent reinforcement learning model of common-pool resource appropriation. In *Advances in Neural Information Processing Systems*, pages 3643–3652.

Rabinowitz, N. C., Perbet, F., Song, H. F., Zhang, C., Eslami, S., and Botvinick, M. (2018). Machine Theory of Mind. *arXiv preprint arXiv:1802.07740*.

Raileanu, R., Denton, E., Szlam, A., and Fergus, R. (2018). Modeling Others using Oneself in Multi-Agent Reinforcement Learning. *arXiv preprint arXiv:1802.09640*.

Schmid, K., Belzner, L., Gabor, T., and Phan, T. (2018). Action markets in deep multi-agent reinforcement learning. In *International Conference on Artificial Neural Networks*, pages 240–249. Springer.

Sutton, R. and Barto, A. (2018). *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning series. MIT Press.

Wang, W., Hao, J., Wang, Y., and Taylor, M. (2018). Towards cooperation in sequential prisoner's dilemmas: a deep multiagent reinforcement learning approach. *arXiv preprint arXiv:1803.00162*.

Winett, R., Kagel, J., Battalio, R., and Winkler, R. (1978). The effects of monetary rebates, feedback, and information on residental energy consumption. *Journal of Applied Psychology*, 63:73–80.

Yang, J., Li, A., Farajtabar, M., Sunehag, P., Hughes, E., and Zha, H. (2020). Learning to incentivize other learning agents. *arXiv preprint arXiv:2006.06051*.