Check for
updates

Investigating the impact of direct punishment on the emergence of cooperation in multi-agent reinforcement learning systems

Nayana Dasgupta[1] · Mirco Musolesi[1,2]

Abstract

Solving the problem of cooperation is fundamentally important for the creation and maintenance of functional societies. Problems of cooperation are omnipresent within human society, with examples ranging from navigating busy road junctions to negotiating treaties. As the use of AI becomes more pervasive throughout society, the need for socially intelligent agents capable of navigating these complex cooperative dilemmas is becoming increasingly evident. Direct punishment is a ubiquitous social mechanism that has been shown to foster the emergence of cooperation in both humans and non-humans. In the natural world, direct punishment is often strongly coupled with partner selection and reputation and used in conjunction with third-party punishment. The interactions between these mechanisms could potentially enhance the emergence of cooperation within populations. However, no previous work has evaluated the learning dynamics and outcomes emerging from multi-agent reinforcement learning populations that combine these mechanisms. This paper addresses this gap. It presents a comprehensive analysis and evaluation of the behaviors and learning dynamics associated with direct punishment, third-party punishment, partner selection, and reputation. Finally, we discuss the implications of using these mechanisms on the design of cooperative AI systems.

✉ Mirco Musolesi
m.musolesi@ucl.ac.uk

Nayana Dasgupta
nayana.dasgupta.18@alumni.ucl.ac.uk

[1] Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK

[2] Department of Computer Science and Engineering, University of Bologna, Via del Risorgimento 2, 40136 Bologna, Italy

⌀ Springer

1 Introduction

The evolution of cooperation has played a pivotal role in the success of the human species, enabling the development of complex societies capable of performing extraordinary feats of collaboration to achieve collectively beneficial outcomes. As such, problems of cooperation occur at all levels of human society, from drivers contemplating whether to give way at busy junctions, to world leaders negotiating carbon reduction targets [1]. Therefore, as the use of artificial intelligence within society becomes more pervasive, the need for socially intelligent agents capable of navigating these complex cooperative dilemmas is becoming increasingly evident [1, 2]. However, the development of socially intelligent agents is challenging, as cooperation is a complex cognitive skill [3]. The factors that influence its evolution are still not well understood [4, 5], despite the intense interest in this problem over the past decades [6–9].

Social mechanisms are powerful tools that can influence the behaviors of populations towards socially responsible choices [1, 10–12]. These mechanisms include punishment [13–15], partner selection [16–21], and reputation [22–25]. Consequently, numerous studies have proposed the use of these social mechanisms within artificial populations to develop cooperative AI systems [1, 5, 26–30].

Although reputation and partner selection can facilitate the emergence of cooperation within small groups, they have a negligible impact on larger populations. Instead, the use of direct and third-party punishment against norm violators has been proposed as a key factor in the emergence of large-scale non-kin cooperation observed within human societies [13, 15, 27, 31, 32]. Direct punishment is particularly interesting in the study of cooperation. Unlike third-party punishment, it has also been observed in non-human populations [13, 32], indicating that direct punishment may be a simpler, yet effective, mechanism for promoting cooperation.

In the natural world, the use of punishment is strongly coupled with the related social mechanisms of partner selection and reputation. Their collective dynamics may aid the emergence of cooperation in societies and incentivize the use of costly punishment to resolve the second-order free rider problem [13, 33]. This combination can also result in competition between agents, which has been associated with the development of unjust punishment of cooperators, potentially detrimental to the emergence of cooperation [34].

Despite the rich connections that exist between these social mechanisms and prior work showing that combinations of other social mechanisms can benefit the emergence of cooperation [29, 35, 36], to the best of our knowledge, no existing work has studied the learning dynamics that emerge from combining direct and third-party punishment with partner selection and reputation in Multi-Agent Reinforcement Learning (MARL) populations. Furthermore, many of the existing studies on punishment do not model the development of unjust punishment or second-order free-riding in their environments [13, 15, 28, 37].

This paper addresses these gaps by presenting a systematic study on the impact of punishment and related social mechanisms on the learning dynamics and behaviors that emerge within populations of MARL agents. It provides an in-depth discussion of the implications of applying these mechanisms to artificial populations. This can guide researchers and practitioners when designing cooperative AI systems that involve social mechanisms. The main contributions of this work are as follows:

- A systematic analysis and evaluation of the learning dynamics and outcomes that emerge from MARL populations using different combinations of direct punishment,

third-party punishment, partner selection, and reputation, including populations that use both direct and third-party punishment.

- An examination of the dynamics of just and unjust punishment that emerge from the competitive environment created by combining punishment, partner selection, and reputation.

- Results showing that direct punishment, when offering a net reward to just punishers, is effective at promoting the emergence of cooperation within populations, and that the effectiveness of direct punishment is improved when it is combined with partner selection and reputation. While third-party punishment leads to a higher proportion of cooperation at equilibrium compared to direct punishment, the combined use of third-party and direct punishment achieves the highest level of societal cooperation in the fewest number of episodes. However, populations using direct punishment achieve significantly higher levels of societal reward, indicating that direct punishment is most effective at maximizing global welfare through cooperation.

2 Background and related work

Cooperation is defined as several self-interested agents performing actions that jointly improve the welfare of an entire group. The natural world is filled with examples of cooperation between self-interested organisms that enable the pursuit of goals beyond the reach of isolated individuals [1]. However, cooperating can often appear illogical for individuals, particularly in situations known as social dilemmas.

Social dilemmas embody conflicts between choices that maximize individual payoffs and those that maximize group payoffs [4, 38]. Each individual's payoff for a non-cooperative choice is higher than their payoff for a cooperative one. However, if too many individuals choose non-cooperative options, a societal collapse occurs, and the entire group receives a lower payoff than if they had all cooperated [3, 39, 40]. This results in the first-order free rider problem, where individuals are motivated to maximize their reward by receiving the group benefit generated by cooperators, while not cooperating themselves [40]. Social dilemmas are ubiquitous within human society, with examples including choosing whether to obey fishing quotas [40], pay taxes [5], or reduce carbon emissions [39]. Therefore, for socially intelligent AI agents to successfully integrate into human society, they must learn to navigate these complex social dilemmas [2].

Social mechanisms are powerful tools for resolving these social dilemmas and guiding populations towards socially responsible choices. Examples of these mechanisms include punishment [15, 27, 28, 30, 35], partner selection [5, 41, 42], reputation [26, 43–45], and commitment [36, 46, 47]. As a result, numerous studies have proposed the integration of social mechanisms within artificial populations for the development of cooperative AI [1, 5, 26–30, 43, 46, 48].

2.1 Punishment

While social mechanisms such as reputation and partner selection have been shown to facilitate the emergence of cooperation within small groups, these mechanisms fail to scale to the size and complexity of the cooperation observed in human populations. Therefore, the use of punishment against norm violators has been proposed as the explanation for the emergence of large-scale non-kin cooperation observed within human societies [13, 15,

27, 31, 32, 49]. Individuals punish by reducing the payoff of another individual, potentially in response to the other individual's behavior. An individual performs direct punishment when they punish their own interaction partner and third-party punishment when they punish an individual who was involved in an interaction with a different individual [13, 50]. Punishment of norm-violators resolves the first-order free rider problem by decreasing the potential payoff of defection and therefore, incentivizing cooperation [15, 31, 34].

While punishment can be beneficial to populations by encouraging the emergence of cooperation, its use incurs costs on both the punisher and the punished [13, 14], ultimately reducing the combined reward and overall productivity of the population [33, 51]. As the use of punishment is costly for the punisher and exposes them to the possibility of retaliation, agents are incentivized to reap the benefits of the cooperative society created by punishers, while avoiding becoming punishers themselves [13, 33, 51]. This second-order free rider problem can result in the breakdown of cooperation within a population, if no one chooses to punish defectors [52].

2.2 Metanorms and centralization

Many studies resolve the second-order free rider problem by applying metanorms [53] that punish those who failed to punish norm-violators [28, 30, 54]. A key limitation of this concept is that infinite regress can occur for example, punishing second-order free riders may result in third-order free riders. This infinite regress can be mitigated by individuals contributing to a centralized fine fund prior to a game, enabling the identification of second order free riders even if there are no first-order free riders [54]. This approach represents the first step towards centralized, institutional punishment that relies on a designated punisher to deliver punishments on behalf of a group [13].

While centralized punishment systems such as a police force are common in modern human societies, for much of human history populations have relied upon decentralized punishment that is delivered voluntarily by individuals. Therefore, the use of decentralized punishment may be more closely related to the emergence of cooperation within early humans [27]. Additionally, decentralized punishment schemes may be more effective in distributed systems such as peer-to-peer networks where agents may not be able to agree on a trusted central punisher or where interactions are high speed and high magnitude [48]. Decentralized punishment systems can be further decomposed into top-down and bottom-up normative orders. A bottom-up normative order allows populations to develop their own norms and sanctioning schemes based on their environment [26, 55, 56]. This is often used to study norm emergence and coordination within populations. Whereas, a top-down normative order does not model the development of social norms within a population and instead models a specific form of law enforcement where a centralized institution creates laws but, enforcement is decentralized and voluntary [27, 28, 57]. This isolates the problem of determining the dynamics of how agents engage in punishment from the problem of coordinating social norms within a population.

Some societies resolve the second-order free rider problem by intrinsically or extrinsically rewarding the use of just punishment, for example by allowing the punishing agent to take resources from the punished agent [27]. Alternatively, the second-order free rider problem can also be overcome by combining the use of punishment with other social mechanisms such as partner selection and reputation [13, 15]. Agents that justly punish defectors can receive reputational gains, as just punishment is a costly signal of social responsibility [51]. Therefore, provided that the benefits of a good reputation for example, access to high

quality interaction partners via reputation-based partner selection, outweigh the costs of punishment, agents may be motivated to punish regardless [13].

2.3 Reputation

An individual's reputation is a record of their prior behaviors. This record can be formed from an agent's own experiences (direct reciprocity) or through the experiences of other agents shared via gossip or other communication mechanisms (indirect reciprocity) [22–25, 43]. The availability of a reputation mechanism allows agents to condition their behaviors towards interaction partners based on their partner's past actions. For example, an agent may choose to only cooperate with agents that have a reputation of prosocial behaviors, to decrease the risk of being exploited by a defector [29]. Studies have shown that this conditional behavior can resolve the first-order free rider problem and extend cooperation beyond dyadic interactions [23] by providing a competitive advantage to cooperators and reducing the potential payoff of defection.

For a population's reputation scheme to be trustworthy, an individual's reputation must be built on costly signals of prosocial intent [16, 58, 59], such as choosing to cooperate despite the risk of exploitation or choosing to actively punish defectors [15, 21, 23, 24, 60]. Therefore, as building a trustworthy reputation is costly, cooperators can only gain a competitive advantage over defectors if the benefits expected from a positive reputation exceed the costs involved in building it.

Direct reciprocity relies on repeated interactions between the same individuals, where each partner provides a benefit to the other [43]. These repeated interactions produce strong pairwise ties that encourage the development of cooperation via conditional behaviors [61]. However, many human interactions, such as donating money to charity, are asymmetric and short-lived [62]. These uni-directional edges also facilitate the emergence of cooperation within networks [63] and are motivated by indirect reciprocity. The latter enables the emergence of cooperation even when interaction partners lack joint histories by aiding the spread of information between unrelated agents within a population [43, 62]. Additionally, indirect reciprocity enables information to spread between independent social structures, such that success in one domain can influence success in other domains [64]. However, indirect reciprocity has substantial cognitive demands as agents must be able to monitor the reputations of others, and so its use is rare in non-human organisms [62].

The simplest form of indirect reciprocity is first-order reciprocity, where an individual's reputation is based solely on their previous actions. Image scoring is a type of first-order reciprocity that increases an agent's reputation when they cooperate and decreases it when they defect [65]. However, image scoring is unstable as individuals have no incentive to defect against defectors to defend themselves from exploitation, because any form of defection would harm their own reputation [43]. In [43], the authors show that it is possible for cooperation to emerge using first-order reciprocity by applying a generous scoring strategy. However, [43] only considers retaliatory defection, thereby ignoring the possibility of specialized punishment mechanisms being used alongside first-order reciprocity strategies such as image scoring. Reputation is often closely linked with punishment, as agents that justly punish defectors often gain a positive reputation within their population. This is because just punishment is a costly signal of social responsibility [51, 58].

In [26], the authors further illustrate that reputation is most effective when it acts in tandem with other mechanisms. They study how populations of independent Q-learning agents learn to use reputations to coordinate their actions in the Iterated Prisoner's

Dilemma. They find that even when environments contain a centralized reputation scheme, which stability predictions demonstrate to be effective for the emergence of cooperation, populations of reinforcement learning (RL) agents fail to achieve cooperation and converge to inefficient equilibria. Therefore, reputation alone is not a strong factor in the emergence of cooperation within populations. Instead, it acts in collaboration with other mechanisms such as punishment [13, 15, 20, 21, 60].

The combination of reputation and punishment can help resolve the second-order free rider problem. Agents with higher reputations enjoy superior access to resources, such as higher quality interaction partners, compared to those with lower reputations [22–25]. Therefore, provided that the benefits of a high reputation outweigh the costs involved in performing just punishment, agents may be motivated to punish due to the reputational gains they may receive [13]. The reputational gains from third-party and direct punishment may differ, as societies are more likely to reward third-party punishers more than direct punishers. This is because third-party punishment is generally seen as a costly act of virtue, while direct punishment is associated with more selfish motivations such as retribution [13, 33, 66].

2.4 Partner selection

Social networks are dynamic; individuals do not always interact with the same set of partners, but instead cut or form ties with specific individuals [21, 23, 67]. Network transitions, regardless of whether they are endogenous or exogenous in nature, can be beneficial for the emergence of cooperation [61]. Partner selection allows individuals to choose who they interact with based on the attributes of potential partners such as their reputation [5]. Reputation plays an important role in enabling cooperation to emerge via partner selection by providing individuals with information about the prior behaviors of potential partners [16–21, 23]. Individuals benefit from choosing high quality, cooperative partners that are unlikely to exploit them and so prefer to select interaction partners with cooperative reputations [17, 22, 31]. Assuming that reputations offer an honest signal of an agent's prosocial intent, cooperators can gain a competitive advantage over defectors by exclusively interacting with other cooperators. The resulting formation of cooperative clusters allows prosocial agents to reap the benefits of mutual cooperation and protect themselves from invasion by defectors [23, 29, 68].

The emergence of cooperative clusters can result in heterogeneous networks where cooperative individuals have far more interaction partners than others. While heterogeneous networks can improve the total wealth of a population, they can also lead to inequalities within societies, such that the least connected individuals may be better off in a state of societal collapse [69]. However, this finding does not consider the possibility of individuals improving their reputations to become a more desirable interaction partner and therefore, more prosperous. Individuals trying to improve their reputation relative to other agents can result in competitive altruism, where individuals compete with each to gain the highest reputation and appear more attractive as an interaction partner [16, 18, 22]. This competition can result in the global emergence of cooperation within populations [31] and incentivize the use of just punishment, therefore aiding the resolution of the second-order free rider problem.

The efficacy of the competitive altruism induced by partner selection in RL populations has been confirmed by multiple studies [5, 70]. In [5], the authors show that allowing populations of DQN agents to select their partners based on their reputation in the

Iterated Prisoner's Dilemma results in the emergence of cooperation, as agents learn to prefer partners with a history of prosocial behavior. However, Anastassacos et al. only use an agent's most recent action to determine their public reputation. This could allow anti-social agents to dishonestly signal prosocial intent by occasionally cooperating to gain a cooperative partner and then exploiting them, however this possibility is not explored. In [70], the authors further confirm that Q-learning agents are able to successfully leverage partner selection in a dynamic environment, illustrating the robustness of partner selection. However, this study does not explicitly study the problem of cooperation and does not consider how reputational information can spread between agents through mechanisms such as gossip, instead relying on private reputations that are only updated with the agent's direct experiences. Together, these results indicate that partner selection is a robust social mechanism that is effective in enabling the emergence of cooperation in artificial populations.

2.5 Unjust punishment

Unjust punishment, also known as anti-social punishment, occurs when punishment is unjustly applied to cooperators. The presence of unjust punishment can be detrimental to the evolution of cooperation as the punishment of cooperators reduces the incentives to cooperate and therefore, encourages pervasive defection [34, 52, 71, 72]. The competitive environment created by combining partner selection, reputation and punishment can encourage the development of unjust punishment as agents may use punishment to diminish the standing of their peers and so become a more desirable interaction partner [34]. However, there is also evidence that unjust punishment can be restrained by commitment mechanisms [35, 47]. Reputation based partner selection can be viewed as an implicit commitment mechanism, such that individuals interact as though they had a prior agreement to prevent mutual risk to their reputations [35, 36, 47]. As a result, an important factor determining the emergence of unjust punishment is the trade-off between the benefits of unjustly punishing competitors with the possibility of reputational losses [33, 34].

Many of the existing studies on punishment [13, 15, 28, 37] have assumed that agents are only able to punish defectors and so do not consider the impact of unjust punishment on social dynamics. For example, in [28], the authors study how punishment enables the evolution of division of labor in populations of RL agents. However, [28] relies on the assumption that a metanorm exists that forces agents to punish according to a centralized social sanctioning matrix, but do not consider how this metanorm should be modeled in the environment. This means the study fails to investigate dynamics associated with use of unjust punishment or second-order free riding.

One of the few studies that model unjust punishment is [27]. In [27], the authors investigate how the use of third-party punishment impacts populations of RL agents within a spatio-temporally extended environment of collecting berries. They show that the frequency of unjust punishment in a population decreases in environments with "silly rules" that do not confer any benefits to individuals or society. The presence of "silly rules" increases the legibility of the norms within a system [73], thus making it easier for agents to learn how to perform just punishment. However, [27] fails to evaluate the impact of third-party punishment on populations experiencing social dilemmas, as berries grow in abundance in the environment used in the study and there is no tension between individual-payoff maximizing and group-payoff maximizing actions. Moreover, Köster et al. use a fixed population size of 12 agents and do not evaluate whether their findings can generalize to larger or smaller populations. These limitations

are somewhat resolved by [57], which investigates the impact of third-party punishment on populations of Q-learning agents playing the Prisoner's Dilemma. They confirm that third-party punishment can successfully promote the emergence of cooperation in populations of up to 128 agents.

2.6 Combinations of social mechanisms

While there is an abundance of literature examining social mechanisms in isolation, there are far fewer studies investigating the dynamics arising from combinations of them. However, the results from these studies show that the dynamics between social mechanisms generally benefit the emergence of cooperation. For example, populations that combine parity, sympathy and reciprocity achieve higher proportions of individual and social welfare compared to populations using a single mechanism [29]; moreover, populations that combine the use of direct punishment with commitment achieve higher proportions of cooperation than if commitment was used alone [35, 47]. Furthermore, the combination of reputation and commitment enables the emergence of cooperation, even in environments without repeated interactions, by targeting cooperation to those who are true to their word [36]. These results reflect the continued co-existence of social mechanisms in the natural world and illustrate the importance of considering interactions between social mechanisms when designing cooperative AI systems.

2.7 Limitations of existing literature

Punishment is a powerful method for shaping the behaviors of a population [15]. Direct punishment is a particularly interesting form of punishment, as unlike third-party punishment, it has also been observed in non-human populations [13, 32]. This indicates that direct punishment may be a simpler but, still effective mechanism for encouraging the development of cooperation within populations. However, there is limited research comparing the dynamics emerging from direct and third-party punishment in MARL populations. Moreover, while there has been extensive research examining direct punishment [51, 66] and third-party punishment separately [27, 32, 33, 55, 57, 74], research into their combined use has been far more limited.

Additionally, both forms of punishment have several limitations that can render them ineffectual. These include the development of unjust punishment [34, 52, 71, 72] or second-order free riding [13, 33, 51]. Moreover, the high cost of punishment can make its use unprofitable for populations, regardless of the proportion of cooperation it produces. In the absence of centralized institutions, these issues could potentially be resolved by combining the use of punishment with reputation and partner selection. Despite the rich connections that exist between punishment and other social mechanisms such as partner selection and reputation, research into the dynamics emerging from their combined use has been limited. To the best of our knowledge, there is no existing literature examining the learning dynamics emerging from the combination of direct and third-party punishment, partner selection and reputation in populations of MARL agents. Moreover, many of the existing studies of punishment fail to model the existence of unjust punishment or second order free-riding in their environments [13, 15, 28, 37].

3 Approach

We aim to provide a comprehensive and systematic analysis of the learning dynamics and behaviors associated with the use of direct and third-party punishment in societies of artificial agents and how these dynamics change when combined with the related social mechanisms of partner selection and reputation. To achieve this, we use a series of simulations where populations play the Iterated Prisoner's Dilemma in the context of several different combinations of social mechanisms. This is done by splitting each simulation into stages, where each of the social mechanisms being studied are associated with a specific interval in the simulation. Each stage can be added or removed to explore different combinations of social mechanisms. This allows for the impact of direct or third-party punishment on the emergence of cooperation within populations to be isolated. Moreover, this enables the combined impact of all the mechanisms to be investigated.

3.1 Iterated Prisoner's Dilemma

The Iterated Prisoner's Dilemma is used to model the cooperative social dilemma that each population must overcome through the use of social mechanisms. This dilemma has been extensively used to study the emergence of cooperation within populations [4, 17, 51, 75]. The Iterated Prisoner's Dilemma involves a pair of agents repeatedly playing the Prisoner's Dilemma, characterized by the payoff matrix in Table 1.
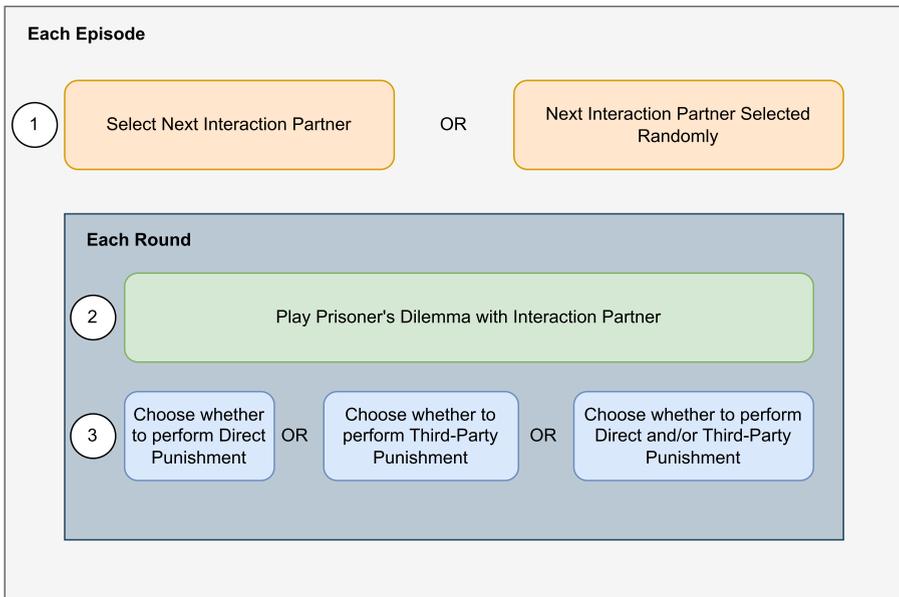
3.2 Simulation

A simulation, illustrated in Fig. 1, is constructed to model the impact of social mechanisms on populations of independent and identically constructed learning agents experiencing the Iterated Prisoner's Dilemma. Each episode in the simulation consists of up to three distinct stages. The first stage in an episode involves agents being paired with their interaction partners, either through partner selection or random matching. Then, for each round every pair of agents will play the Prisoner's Dilemma with their partner and then immediately choose whether or not to punish another agent, which may or may not be their partner, based on the decisions made in the previous game. Each simulation consists of 2000 episodes, with ten rounds per episode.

Table 1 Payoff matrix for the Iterated Prisoner's Dilemma

			Player 2	
🟥	Reward		C	D
🟧	Sucker			
🟦	Temptation	*Player 1*	C	3, 3	0, 4
🟥	Punishment		D	4, 0	1, 1

Fig. 1 Each episode in a simulation consists of up to three distinct stages. In the first stage, depending on whether partner selection is being studied in the current simulation, agents either select their next interaction partner using their partner selection DQN model or their partner is selected randomly out of all other agents in the population. Agents then play the Prisoner's Dilemma with their partner in the second stage before, choosing whether or not to carry out punishment in the third stage. The third stage can consist of agents performing direct punishment, third-party punishment or both direct and third-party punishment depending on the combination of social mechanisms being studied in the current simulation. The second and third stages repeat consecutively for each round in the episode, while the first stage occurs only once at the start of an episode

3.2.1 Reputation

Each agent in the population is assigned a scalar integer reputation that encodes the societal value of their actions. This reputational information is updated with their playing and punishing actions throughout each agent's lifetime. This model of reputation is a more costly and therefore, trustworthy signal of prosocial behavior compared to previous studies of MARL populations that only used an agent's previous playing action to determine their reputation [5]. Reputational information can be included in the partner selection, dilemma game playing or punishment states to inform an agent's decision making process within the corresponding stages of the simulation. This mimics how reputational information is used in conjunction with other social mechanisms within the natural world. Every agent's reputation is publicly available to all agents in the population, using the assumption that this information is shared using a transfer mechanism such as gossip and that this transfer mechanism is complete and honest. This simplifying assumption has been used by much of the existing literature, as it isolates the impact of reputation on the emergence of cooperation from the complexity of agents learning to assign reputations to others [26].

While agents can use both reputations and their experience of past interactions to condition their behaviors towards other agents, experiments detailed in Appendix D show that this is not sufficient to promote the emergence of cooperation within MARL populations.

This provides further evidence to the idea that reputational information is not by itself a strong factor in the emergence of cooperation, but rather acts in tandem with other mechanisms. As a result, this study will consider reputation as an auxiliary social mechanism that facilitates the use of others within populations.

3.2.2 Stage 1: Partner selection

At the start of every episode, every agent is paired with another agent in the population to be their interaction partner for the next ten rounds. Agents cannot be paired with themselves and all agents can both select and be selected in an episode. There are no restrictions on how many times an agent can be paired with other agents in the same episode.

If the combination of social mechanisms being investigated by a simulation includes partner selection, then at the start of every episode, every agent, regardless of their current reputation, is able to select their interaction partner for the episode using the reputations of the other agents in the population. The selected agent cannot refuse to play with the agent that selected them. Therefore, every agent will play at least ten rounds with their selected interaction partner. However, some agents will play more if they are selected by other agents in the population. It is up to each agent to learn how to use the reputational information available to select a valuable interaction partner. Therefore, if enough agents learn a coordinated view on the reputations they prefer their interaction partners to have, those socially desirable agents will gain a competitive advantage as they will be selected and therefore, play more often. This creates normative pressure towards the behaviors that foster the reputation deemed desirable by the population and models the reputation-based partner selection observed within the natural world [16, 21, 22]. If partner selection is not included in the combination of social mechanisms being investigated by a simulation, every agent is randomly paired with another agent in the population.

3.2.3 Stage 2: Prisoner's Dilemma

Each round of the episode begins with each of the paired agents playing an iteration of the Prisoner's Dilemma. This stage is an essential part of the simulation, as it models the main social dilemma that the population must overcome. As the choice to cooperate in a social dilemma is a costly signal of social responsibility [5, 23, 25], cooperating increases the value of an agent's reputation by one while defection decreases it by one. The rewards received by each agent playing the Prisoner's Dilemma are specified by the payoff matrix in Table 1.

3.2.4 Stage 3: Punishment

After each pair of agents has played an iteration of the Prisoner's Dilemma, the next step in each round is the application of punishment to agents within the population. This punishment stage involves each agent choosing whether or not to punish a target agent based on the actions that the target agent took within the previous Prisoner's Dilemma game. Depending on the combination of social mechanisms being studied in the simulation, the punishment stage can consist of either third-party punishment, direct punishment or both third-party and direct punishment.

Third-party punishment involves an agent that is unrelated to an interaction choosing whether to punish an agent who is directly involved in the interaction [13, 32]. Therefore,

for each interaction between a pair of agents *A* and *B* in a simulation that contains third-party punishment, two agents *P* and *K* that are not involved with the interaction between *A* and *B* are randomly selected to act as potential third-party punishers. Agent *P* decides whether to punish agent *A* and agent *K* decides whether to punish agent *B*. Instead, direct punishment involves the individuals involved in the interaction deciding whether to punish each other after interacting [50]. Therefore, given an interaction between a pair of agents *A* and *B* in a simulation that contains direct punishment, agent *A* decides whether to punish agent *B* and vice versa.

In the natural world, the decision to apply punishment is heavily related to the reputational benefits punishers receive from delivering punishments to others [13, 15]. However, while just punishment of defectors is looked upon favorably by society [13, 15, 33], unjust punishment of cooperators is perceived unfavorably [34, 72]. To mirror this behavior, the simulation rewards agents who justly punish a defector by increasing their reputation by two and decreases the reputation of agents by three when they unjustly punish a cooperator.

As performing punishment is costly for the punisher in the natural world [13, 14], the decision to perform punishment in the simulation reduces the reward of the punishing agent by ten and the reduces the reward of the punished agent by three. Two different reward schemes for just punishment were evaluated. In the first reward scheme, referred to as Scheme 1, if the punisher performs just punishment of a defector, they gain seven units of reward resulting in the punisher suffering a net loss of 3 units of reward. This reward scheme relies on the presence of partner selection and reputation to resolve the second order free rider problem. In the second reward scheme, referred to as Scheme 2, if the punisher performs just punishment of a defector, they gain twelve units of reward. Therefore, just punishment results in the punisher receiving a net profit of two units of reward. This reward scheme was inspired by [27] and models either intrinsic or extrinsic rewards for just punishment. These rewards create an incentive for agents to learn to perform just punishment and reflect a common judicial system throughout human history where a centralized institution labels antisocial behavior but enforcement is decentralized and rewarded by being able to take the transgressor's property [27].

In each of the experiments, the population consisted of five agents. As outlined in Appendix C, additional experiments were carried out to determine whether cooperation behaviors changed with different population sizes and these experiments showed that increasing the number of agents did not have a strong effect on the general behaviors observed and the slight differences between population sizes are reported upon and analyzed in the Appendix.

3.3 Learning framework

In order to create artificial populations of agents that are capable of simulating responses to mixed-motive cooperative dilemmas, each agent is modeled as an independent actor [5, 57, 76, 77]. This allows agents to have goals that are not fully aligned with others in the population and allows agents to avoid sharing all of their information e.g. reward signals or observations with others. This model may also contribute to current work modeling human-human and AI-human interactions, as humans are typically unable to share their reward signals or observations with other agents [55].

Populations consist of *N* identically constructed learning agents. Each learning agent consists of up to three independent Deep Q-Network (DQN) models, each specialized for a different agent ability. These agent abilities are the following: choosing an interaction

partner, playing the Prisoner's Dilemma and choosing whether to punish. This model extends the approach proposed by [5] to include an additional DQN model specialized for punishment. This multi-model design allows for greater flexibility compared to a single model design, as different hyper-parameters can be used for each DQN model to optimize the performance of each agent ability. This design also reduces the complexity of the neural networks needed for each DQN model. Each learning agent in the population will only contain a DQN model for a specific ability if it is required by the social mechanism combination being investigated in the current experiment.

3.3.1 Deep Q-network learning

RL algorithms such as Deep Q-Network (DQN) allow agents to continuously learn from their experiences and organically discover strategies [26]. This results in a highly dynamic environment where independent agents are learning simultaneously and where every agent behavior needs to be learned from stimuli, allowing normative legibility to be modeled via how easy it is for agents to learn a behavior. The RL learning dynamics also result in temporal dependencies between the learned behaviors: for example, agents might need to learn to punish defectors prior to learning to cooperate [27]. Therefore, the use of DQN, as a RL algorithm, allows the learning dynamics that emerge from populations using social dilemmas to be made transparent. DQN was selected as the most parsimonious approach, which is capable of dealing with large state spaces and can be scaled up to potentially very large populations at the same time.

Agents are independently trained using Q-learning [78] to learn an estimate of the optimal action-value function $Q_\pi(s, a) : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ and therefore, the optimal policy π. With DQN, $Q_\pi(s, a)$ is parameterized by a neural network. After each interaction, each agent updates $Q_\pi(s, a)$ using Eq. 1, where s is the state they observed, a is the action they took, r is the reward received and s' is the next state. To determine which action should be taken in each state, each agent applies an ϵ-greedy policy.

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left[r + \gamma \max_{a' \in \mathcal{A}} Q(s', a') - Q(s, a) \right] \tag{1}$$

3.3.2 Model

To formalize, the simulation is a N player Markov Decision Process. When partner selection is used in a population, agents observe a state $S_{select} \in \mathbb{R}^N$ consisting of an array containing all agent reputations within the population. The partner selection DQN model then outputs the Q-values of each potential interaction partner, with the partner that has the maximum Q-value being chosen as the agent's interaction partner.

During the Prisoner's Dilemma stage, the input states S_{play} observed by populations using direct punishment and populations using third-party punishment or combined direct and third party punishment differ. Experiments, detailed in Appendix B, show that including reputational information in S_{play} is beneficial to the emergence of cooperation when a population adopts third-party punishment, while this is not the case for direct punishment. Therefore, to ensure that no combination was unfairly advantaged, the optimal input state S_{play} was used for each social mechanism combination. Therefore, populations using third-party punishment or both third-party and direct punishment observe a state $S_{play} \in \mathbb{R}^4$ that consists of both their own and their partner's reputation and previous playing action. While

populations using direct punishment observe a state $S_{play} \in \mathbb{R}^2$ that consists of both their own and their partner's previous playing action. The dilemma game playing DQN model then outputs the Q-values associated with cooperating (0) and defecting (1), with the action that has the maximum Q-value being selected.

During the punishment stage, agents observe a state $S_{punish} \in \mathbb{R}^2$ consisting of the previous playing actions of both the agents involved in the interaction being judged. The punishment DQN model then outputs the Q-values associated with the agent not punishing (0) or punishing (1), with the action that has the maximum Q-value being selected. A series of experiments, outlined in Appendix B, found that adding reputational information to the punishment state negatively impacted the emergence of cooperation within populations for both forms of punishment, and so reputational information was not included in the punishment state. In the case of populations using both third-party and direct punishment, the same punishment DQN model is used for both types.

Figure 2 illustrates the network architectures of each of the DQN models associated with an agent ability. Each DQN model is parameterized using a single hidden layer, with 128 neurons and a ReLU activation function. The use of 128 neurons resulted in more consistent results compared to smaller networks over multiple runs.

The hyper-parameters for each DQN model are optimized for their specific task and they are shared across all agents. A linearly decaying ϵ-greedy policy (where a maximum ϵ value decays to a fixed minimum ϵ value) was used alongside the following hyper-parameters: the partner selection, dilemma game playing and punishment models all shared a maximum ϵ of 0.8889, a discount rate of 0.9, a batch size of 100 and updated the weights of their target network to match their online network every 200 steps. The partner selection model used a minimum ϵ of 0.0001 and a learning rate of 0.01, while the playing model used a minimum ϵ of 0.01 and a learning rate of 0.1. Instead, the punishment model used a minimum ϵ of 0.2 and a learning rate of 0.001. A description of the experiments used to select these hyper-parameters is reported in the Appendix.

4 Results

This section presents a systematic analysis on the impact of direct punishment on populations of MARL agents, illustrating that direct punishment has a positive impact on the emergence of cooperation. Additionally, populations that combine the use of direct punishment with reputation and partner selection experience further gains in societal cooperation. The results show that populations using third-party punishment consistently achieve higher levels of cooperation at convergence compared to populations using direct punishment. At the same time, populations that use both third-party and direct punishment achieve the highest level of societal cooperation in the least number of episodes. Additionally, direct punishment achieves the highest level of societal reward at convergence, indicating that it is most effective at maximizing global welfare through cooperation.

4.1 Description of the experiments

Each experiment involved 2000 episodes, with each episode consisting of ten rounds. Experimental trials showed that 2000 episodes was sufficient for convergence. The experiments were each repeated twenty times. The experiments evaluated the impact
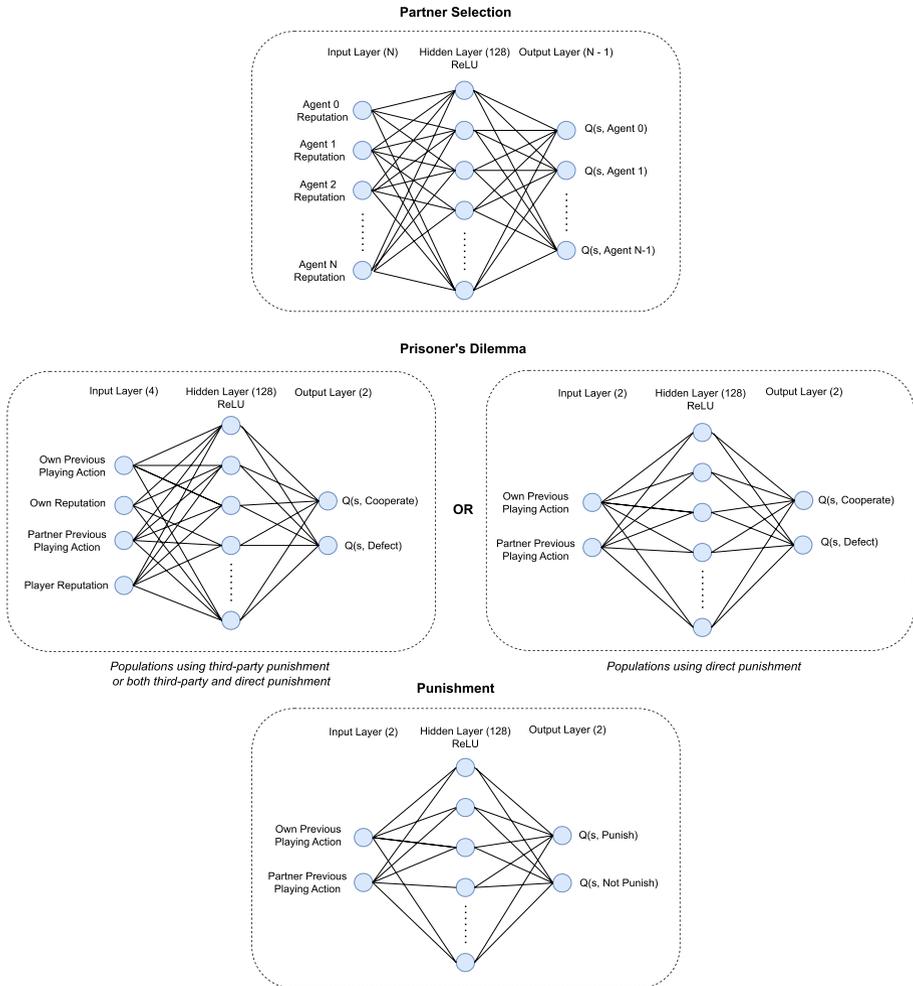
Fig. 2 Each agent consists of up to three independent DQN models, each specialized for a different agent ability

of the following combinations of social mechanisms on the emergence of cooperation within MARL populations:

- *TPP-S* third-party punishment with partner selection.
- *TPP* third-party punishment.
- *DP-S* direct punishment with partner selection.
- *DP* direct punishment.
- *TPPDP-S* third-party punishment with direct punishment and partner selection.
- *TPPDP* third-party punishment with direct punishment.

4.2 Evaluation metrics

Several metrics were designed to comprehensively evaluate the learning dynamics and behaviors displayed within simulations of cooperative social dilemmas. The rolling mean of each metric and the associated 95% confidence interval across twenty repeats was calculated, with a rolling window of 100 episodes. In particular, we considered the following metrics:

Cooperation per episode Percentage of cooperative actions taken by the entire population in the dilemma game, in each episode.

Cooperator selections per episode Percentage of agents selected as partners per episode who performed cooperative actions in the dilemma game for the majority of the previous episode.

Punishment per episode Percentage of punishing actions taken by the entire population in each episode.

Percentage of selected punishers per episode Percentage of agents selected as partners per episode who punished in the previous episode.

Ratio of just punishment to unjust punishment per episode Percentage of punishing actions taken by the entire population that were applied to defectors in each episode.

Just punisher selections per episode Percentage of agents selected as partners per episode who performed just punishment of defectors for the majority of the punishment opportunities they had in the previous episode.

Societal reward per episode Total reward obtained by the entire population per episode.

Societal reputation per episode Total reputation obtained by the entire population per episode, where reputation is cumulative over episodes.

4.3 Comparing just punishment reward schemes

Figure 3a shows that populations relying on Scheme 2, where additional intrinsic or extrinsic rewards are associated with just punishment, successfully converge to cooperation in all cases. However, when populations use Scheme 1 to reward just punishment, all populations

(a) Cooperation   (b) Societal Reward
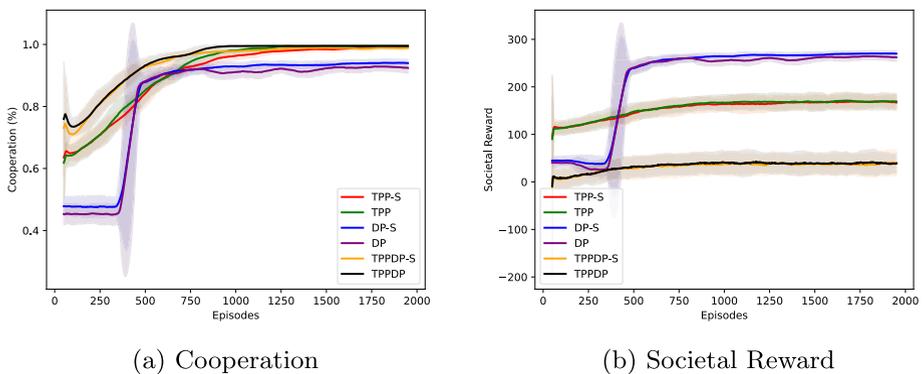
Fig. 3 Populations using direct punishment or direct punishment with partner selection learn to cooperate, but converge to a lower proportion of cooperation compared to populations using third-party punishment or both direct and third-party punishment. Despite this, populations using direct punishment achieve significantly higher levels of societal reward at convergence. This indicates that populations combining direct punishment with partner selection are most effective at maximizing global welfare through cooperation

fail to converge to cooperation, as they fail to learn to apply punishment. This indicates that the presence of partner selection and reputation is not sufficient to resolve the second-order free rider problem in MARL populations. The figures illustrating the social dynamics that emerge from the use of Scheme 1 can be found in Section E of the Appendix. All further results rely on Scheme 2, with the assumption that there are additional intrinsic or extrinsic rewards associated with just punishment within the environment.

4.4 Measuring the success of direct punishment

The success of multi-agent populations can be measured by both their proportion of prosocial behavior and the population's combined reward. Figure 3a shows that direct punishment is effective at encouraging the emergence of cooperation within populations. However, populations using direct punishment or direct punishment with partner selection achieved slightly lower proportions of cooperation compared to populations using third-party punishment. This supports previous findings that direct punishment is less effective at encouraging the development of cooperative behavior in populations compared to third-party punishment [13, 15, 31, 32]. Figure 3a also shows that populations combining direct and third-party punishment not only achieve the highest proportions of cooperation but also converge the fastest. This observation supports the continued existence of both direct and third-party punishment within human societies.

Although direct punishment bears the same costs as the more effective third-party form of punishment, Fig. 3b shows that populations using direct punishment with partner selection achieve the highest levels of societal reward at convergence. These populations are followed closely by those using direct punishment alone. This suggests that societies using direct punishment are more efficient than their counterparts that wield third-party punishment. Therefore, considering that the success of a population is determined by both the proportion of cooperation achieved and overall societal wealth, direct punishment, particularly when combined with partner selection, emerges as the most effective social mechanism for overall global welfare and cooperative behavior.

4.5 Dynamics of just and unjust punishment

To understand the reasons behind the comparative success of populations using direct punishment or direct punishment with partner selection compared to other social mechanisms, we must investigate the dynamics of just and unjust punishment associated with each social mechanism combination.

4.5.1 Direct punishment

Figure 3a shows that prior to the 350th episode, high levels of defection were relatively stable in populations using direct punishment or direct punishment with partner selection. During the same period, the use of punishment in populations using direct punishment or direct punishment with partner selection was the highest of all social mechanism combinations. Additionally, prior to the 185th episode, the populations' have the lowest proportion of just punishment. Therefore, populations using direct punishment or direct punishment with partner selection experience an early period of pervasive unjust punishment. The presence of unjust punishment generates normative pressure that dissuades agents from learning to cooperate. Figures 3b and 4a show that the high levels of defection and unjust

(a) Societal Reputation   (b) Ratio of Just to Unjust Punishment
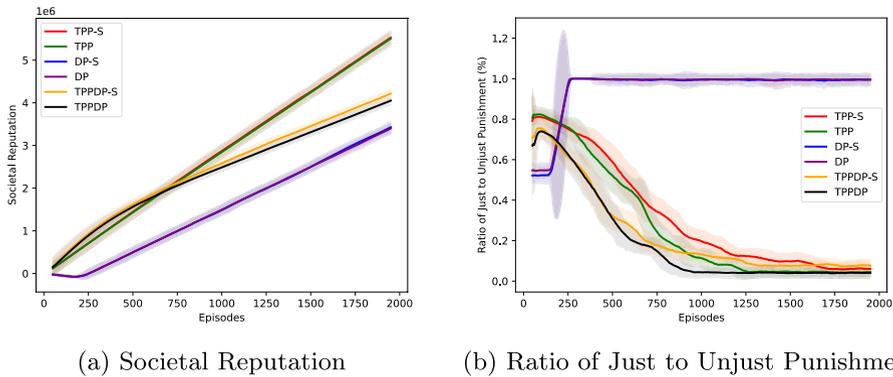
Fig. 4 Populations using direct punishment experience an early period of pervasive unjust punishment prior to the 140th episode, resulting in a decrease in societal reputation. As populations begin to learn to perform just punishment and cooperate, the societal reputation of populations using direct punishment increases but, to a lesser extent compared to populations using third-party punishment or combined third-party and direct punishment

punishment during this period lead to a decline in societal rewards and reputation. The lack of rewards associated with unjust punishment also result in the proportion of punishers within the population remaining stagnant until the 160th episode.

The use of just punishment begins to rise after the 140th episode, as agents rapidly begin to learn that performing just punishment in a population with high levels of defection is highly profitable. Therefore, populations using direct punishment or direct punishment with partner selection learn to always perform just punishment over unjust punishment when they choose to punish. This rapid learning may emerge from the high levels of defection experienced by populations employing direct punishment, as this ensures that many opportunities for just punishment are available. Figure 3a shows that this results in a rapid emergence of cooperation within populations using direct punishment or direct punishment with partner selection at the 350th episode. Figure 3b shows that this leads to a substantial increase in societal reward. However, as the number of defectors within the populations decrease, opportunities for just punishment also decrease. Therefore, as the populations have been trained to avoid unjust punishment, Fig. 5 shows that the overall proportion of punishment decreases to the lowest levels of all the social mechanism combinations by the 475th episode. This reduces the normative pressure for defecting agents to learn to cooperate, resulting in populations using direct punishment or direct punishment with partner selection plateauing to a lower proportion of cooperation compared to other social mechanism combinations.

Figure 5 also shows that the punishment per episode metric is almost inversely proportional to cooperation per episode for every social mechanism combination. This is because the lower levels of defection reduce the availability of profitable just punishing opportunities. As unjust punishment is extremely costly, agents learn to reduce their application of punishment as rates of cooperation increase within a population.

The increased levels of cooperation and just punishment at convergence lead to increases in societal reputation however, Fig. 4a shows that populations using direct punishment or direct punishment with partner selection still lag behind the other social mechanisms at convergence as a result of the comparatively lower levels of cooperation. The stabilization of cooperation levels also leads to a plateauing of societal reward; however, as the overall
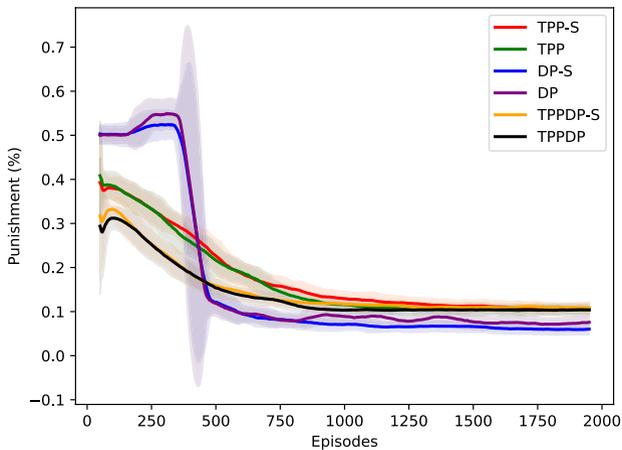
Fig. 5 Punishment per episode. Populations using direct punishment initially have the highest levels of punishment, despite also having the lowest levels of just punishment. As agents begin to learn how to punish justly, the proportion of punishment in the population increases as the act of punishing becomes rewarding for agents. As the levels of cooperation increase within a population, the use of punishment decreases regardless of the social mechanisms used as there are fewer opportunities for profitable just punishment. Populations using direct punishment have the lowest levels of punishment at convergence, resulting in reduced normative pressure to cooperate and therefore, lower levels of cooperation overall

levels of punishment are the lowest of all the social mechanism combinations, the populations achieve a substantially higher level of societal reward at convergence than the other social mechanisms.

4.5.2 Third-party punishment

Populations utilizing third-party punishment or a combination of third-party punishment and partner selection experience different dynamics. Figure 4b shows that prior to the 210th episode, populations using third-party punishment or third-party punishment with partner selection hold the highest ratio of just punishment to unjust punishment. This results in normative pressure to cooperate being applied to agents at an early stage within the simulation, enabling rapid convergence to a high level of cooperation. Consequently, Fig. 3b shows that populations using third-party punishment or third-party punishment with partner selection achieve the highest societal reward of all the social mechanisms until the 412th episode.

However, Fig. 4b shows that populations using third-party punishment (or third-party punishment with partner selection) fail to eradicate unjust punishment. The ratio of just punishment to unjust punishment decreases as cooperation increases. The use of punishment also decreases as opportunities for profitable just punishment decrease. However, as unjust punishment of cooperators still occurs at convergence, the overall levels of punishment within the population remain higher than in populations using direct punishment, despite higher cooperation levels. This suggests that populations using third-party punishment are slower to learn how to punish justly and this may be because the rapid rise in cooperation within the populations means that agents fail to observe a sufficient number of examples of positively rewarding just punishment on defectors for the entire population to learn the behavior.

This limits further increases in societal reward due to the high costs involved with punishment, particularly its unjust form. Nevertheless, populations using third-party punishment or third-party punishment with partner selection achieve the highest societal reputation at convergence. This is as a result of the populations attaining higher levels of cooperation at convergence compared to populations using direct punishment or direct punishment with partner selection, in addition to having lower levels of unjust punishment than populations that combine third-party and direct punishment as the latter has double the number of punishment opportunities.

4.5.3 Combining third-party and direct punishment

Populations applying both third-party and direct punishment experience similar dynamics of just and unjust punishment to populations using third-party punishment. Populations apply just punishment within the first 100 episodes, resulting in the most rapid convergence to the highest level of cooperation of all the social mechanism combinations studied. Similarly to third-party punishment, populations combining third-party and direct punishment fail to eradicate unjust punishment, resulting in higher levels of punishment at convergence compared to populations using direct punishment.

Figure 3b shows that populations that combine direct and third-party punishment always attain the lowest amount of societal reward. This is because these populations have two opportunities for punishment with every interaction, doubling the costs associated with unjust punishment. Similarly, Fig. 4a shows that despite the high levels of cooperation enabling populations combining direct and third-party punishment to have the highest societal reputation prior to the 690th episode, after this point the levels of unjust punishment from the two punishment opportunities lead the populations to lag behind third-party punishment and third-party punishment with partner selection.

4.6 Dynamics of combining direct and third-party punishment

Figure 3a shows that populations using both direct and third-party punishment are the quickest to converge and achieve the highest levels of cooperation. This suggests that direct and third-party punishment work better in tandem and their involvement in the evolution of cooperation may be inter-related. This may be as a result of the increased opportunities for punishment increasing the legibility of the social norms [73] and, therefore, enabling faster learning of cooperative behaviors. The same levels of overall punishment and just punishment were observed between the third-party and direct punishment mechanisms. This is as a result of both mechanisms being controlled by a single punishment model and future research may investigate the impact of splitting the model into further specialized direct and third-party punishment models. Figure 7b further shows that when both third-party and direct punishment are possible within a population, agents are more likely to select just third-party punishers compared to just direct ones. This indicates that third-party punishers are more valued than direct ones and so more influential within the population, providing a justification for the similar dynamics observed between populations using third-party punishment alone and populations using both third-party and direct punishment.

## 4.7 Understanding the impact of partner selection

Figure 6 shows that all populations, regardless of social mechanisms used, learn to select cooperators in nearly all interactions. The populations achieve the same proportion of cooperator selections at convergence, despite differing levels of cooperation at convergence. This indicates that all populations learn to value cooperators, enabling them to gain a competitive advantage by being selected as interaction partners more frequently. The early stages of Fig. 6 are mirrored by levels of cooperation in Fig. 3a as higher levels of cooperation increase the likelihood of selecting a cooperative agent while populations are still in the process of learning to select cooperators. Furthermore, this result also shows that populations are able to discern how to use reputational information to influence their behaviors.

The behavior of punisher selections in Fig. 7a closely follow the levels of punishment in Fig. 5. This indicates that populations, regardless of the social mechanisms used, do not learn to favor punishers over non-punishers. Whereas, the behavior of just punisher selections in Fig. 7b differ. Figure 7b shows that populations using direct punishment with partner selection initially have the lowest level of just partner selections. This reflects the high levels of unjust punishment within the populations. The proportion of just punishment selections increase from the 145th episode, aligning with the rapid rise in the use of just punishment within the population. The percentage of just punisher selections reaches a plateau at the 265th episode, which matches the point where the population learns to always perform just punishment in Fig. 4b.

Following the plateau, the proportion of just punisher selections decreases as the proportion of punishment occurring in the population decreases. However, populations combining direct punishment with partner selection converge to the highest level of just punisher selections of all the social mechanism combinations. This increases the value of performing just punishment by enabling just punishers to participate in more interactions, contributing to the population rapidly learning to perform just punishment in nearly all cases and increasing the normative pressure to cooperate. By preferring cooperators and just punishers, populations that combine direct punishment and partner selection experience higher levels of cooperation at convergence compared to populations that use direct punishment alone. Even prior to the 350th episode and the emergence of just punishment,

Fig. 6 Cooperator selections per episode. All populations, regardless of the social mechanisms used, learn to select cooperators in nearly all interactions at convergence

(a) Punisher Selections   (b) Just Punisher Selections
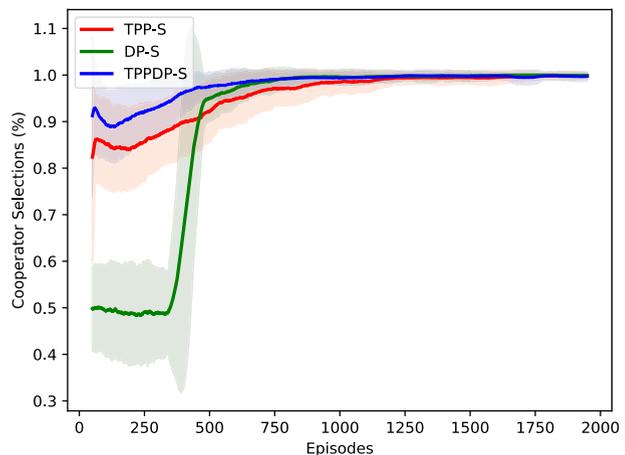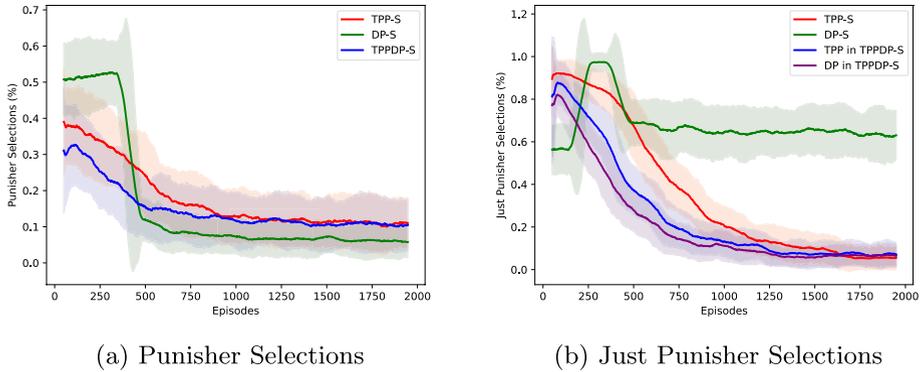
Fig. 7 Fig. 7a shows that populations, regardless of the social mechanism used, do not learn to favor generic punishers over non-punishers. However, Fig. 7b shows that populations using direct punishment do learn to favor just punishers. Populations using direct punishment initially have the lowest levels of just punisher selections, before it rapidly increases at the 145th episode and converges to the highest level of just punishment. This increase in just punisher selections contributes to the rapid learning of just punishment within populations using direct punishment

populations that combined direct punishment with partner selection achieved a higher level of cooperation compared to populations that used direct punishment alone, potentially due to the additional normative pressure of cooperative selections.

Both populations using third-party punishment with partner selection, and populations that combine third-party punishment with direct punishment and partner selection, start with a high level of just punisher selections within their population. This drives the increased levels of just punishment in populations with partner selection, as visible in Fig. 4b. However, as the levels of just punishment decrease within the population, the proportion of just punisher selections also decreases. Figure 3a shows that the inclusion of partner selection in populations using third-party punishment, or a combination of third-party and direct punishment, results in slightly slower convergence to cooperation. This slower convergence suggests that the combination of third-party punishment and partner selection is more complex for agents to learn, leading to a greater number of mistakes during the learning process.

Additionally, Fig. 3a shows that populations using direct punishment without partner selection experience minor fluctuations in cooperation per episode after the 750th episode. These are accompanied by reversed fluctuations in Fig. 5. This suggests that populations applying direct punishment without partner selection experience temporary increases in cooperation, followed by increases in free-riding. This results in cycles of defection and punishment. As these fluctuations are not present in populations that combine direct punishment with partner selection, it can be inferred that the latter may have a smoothing effect on the cooperation achieved by populations utilizing direct punishment.

5 Conclusion and outlook

This study demonstrates that while direct punishment is effective at promoting the emergence of cooperation within populations, those using third-party punishment achieve higher levels of cooperation upon convergence. This provides further evidence that

third-party punishment is the primary driving force for the evolution of large-scale cooperation within non-kin human societies [13, 15, 27, 31].

The comparatively poorer performance of populations using direct punishment may be due to their higher levels of unjust punishment at the start of the simulation. This incentivized pervasive defection that could not be fully resolved before the population started reducing punishment as cooperation increased and the benefits of performing punishment reduced. However, populations that use direct punishment achieve the highest levels of societal reward, indicating that the use of direct punishment results in the most efficient societies. This efficiency may contribute to why direct punishment is preferred in simpler and more resource-constrained non-human populations. This work also shows that populations that combine the use of third-party and direct punishment achieve the highest levels of cooperation overall. This could be a result of the increased number of punishment opportunities, which enhances the legibility of social norms [73], thereby enabling faster learning of cooperative behaviors. This finding reflects the continued presence of both direct and third-party punishment within societies. While the combination of direct punishment with partner selection and reputation led to higher levels of cooperation, populations that used third-party punishment or a combination of third-party and direct punishment were slower to converge when partner selection was introduced. This suggests that the complexity involved in learning how to use combinations of social mechanisms varies between direct and third-party punishment.

It is possible to identify a series of limitations of this work. Firstly, we assume that all agents have access to complete and accurate global reputations, and that all events contributing to an agent's reputation maintain the same level of importance throughout its lifetime. Furthermore, the environment provides a top-down normative order, defining what constitutes just and unjust punishment. Although this simplified model is suitable for the objectives of this work, since it allows for the isolation and analysis of behaviors and learning dynamics emerging from these norms, a more realistic model would involve the agents developing their own social norms. Future research could also consider the emergence of normative orders that allow agents to develop and enforce their own social norms. Additionally, it could investigate how agents adapt to changes in payoffs and norms during the learning process.

This work provides a systematic analysis of several fundamental social mechanisms within MARL systems that can act as a guide for researchers and practitioners when selecting or combining these design dimensions within cooperative AI systems. Therefore, this work not only introduces a new perspective on the role of direct punishment within the evolution of cooperation, but also represents a strong foundation for the creation of new punishment-based techniques as a basis for the design of cooperative AI systems.

Appendix A: Hyperparameter tuning

Each experiment involved 2000 episodes, with each episode consisting of ten rounds. The experiments were each repeated twenty times. Hyperparameter tuning involved a random search of 100 hyper-parameter combinations to find the hyper-parameters that maximized the mean joint reward for all agents over all the repeats. The following hyper-parameter ranges were investigated during the hyper-parameter tuning process (Table 2):

- Maximum Buffer Size $\in [2^x \mid x \in [11, 21)]$

Table 2 Hyperparameters used for all experiments, with a target update of 200, a discount rate of 0.9 and a batch size of 100

	Max Buffer Size	Min ϵ	Max ϵ	ϵ Decay	Learning Rate
Selection Model	131072	0.0001	0.8889	0.3	0.01
Playing Model	131072	0.01	0.8889	0.3	0.1
Punishing Model	524288	0.2	0.8889	0.5	0.001

- Batch Size $\in [2^x \mid x \in [10, 19)]$
- Target Update $\in \{2^x \mid x \in [500, 5001), x\%500 = 0\}$
- Minimum $\epsilon = $ np.linspace(1e-4, 1, num=10)
- Maximum $\epsilon = $ np.linspace(1e-4, 1, num=10)
- ϵ Decay $\in$ np.linspace(1e-4, 0.9, num=10)
- Discount Rate $(\gamma) \in [0.8, 0.9, 0.99]$
- Learning Rate $\in [0.001, 0.01, 0.1]$

Appendix B: Determining optimal reputation and state information composition

Several experiments were carried out to determine what information about an agent's past behaviors should contribute to the calculation of their reputation and how this reputational information should be used, in order to maximize the emergence of cooperation within a population. The results of these experiments provide insights on how providing populations with several varieties of long-term playing and punishing information impacts population dynamics and the emergence of cooperation.

Overview

The first set of experiments involved comparing the levels of cooperation achieved within populations when reputation is calculated using playing behavior alone, punishing behavior alone or both playing and punishing behavior. Another set of experiments evaluated the impact of allowing agents to observe reputational information during playing and punishing decisions, in addition to using reputational information during partner selection. These experiments compared the levels of cooperation achieved by populations when reputational information was added to either the playing state, the punishing state, both the playing and punishing state or neither state. These experiments were conducted on populations that used third-party punishment with partner selection and reputation, as well as populations that used direct punishment with partner selection and reputation.

The following experimental results determine the optimal set of playing and punishing information that should be included in the calculation of agent reputations to maximize the emergence of cooperation within populations. These results also provide an insight on the relative importance of including playing and punishing information within reputations and the usefulness of including reputational information in playing and punishing states.

Populations using third-party punishment with partner selection and reputation

As shown in Fig. 8, cooperation per episode in populations using third-party punishment with partner selection and reputation is maximized when each agent's reputation is calculated using both their playing and punishing behavior. While calculating an agent's reputation using their playing behaviors alone results in a similar outcome to calculating their reputation using both their playing and punishing behaviors, calculating an agent's reputation using their punishment behaviors alone results in substantially lower levels of cooperation at convergence. This indicates that the presence of long-term playing behavior information within the calculation of agent reputations increases the likelihood of the emergence of cooperation within a population, compared to the presence of long-term information about punishing decisions.

The relative importance of including playing and punishing behavior in the calculation of reputation varies during the initial stages of learning. Between the first and the 500th episode, populations calculating agent reputations using punishment behavior alone achieve the highest level of cooperation per episode, with the populations using other forms of reputation achieving significantly lower levels of cooperation per episode. This indicates that during the early stages of learning, information about an agent's punishment behaviors is a more effective signal of their trustworthiness than information about the agent's playing behaviors. Additionally, during the initial stages of learning, populations that calculate an agent's reputation solely based on the agent's playing behaviors achieve slightly higher levels of cooperation per episode compared to populations that calculate an agent's reputation based on both their playing and punishing behaviors. The delayed onset of cooperation associated to the calculation of reputations using both playing and punishing behavior may be due to the increased complexity of learning how to interpret a reputation calculated from two information sources compared to a single information source.
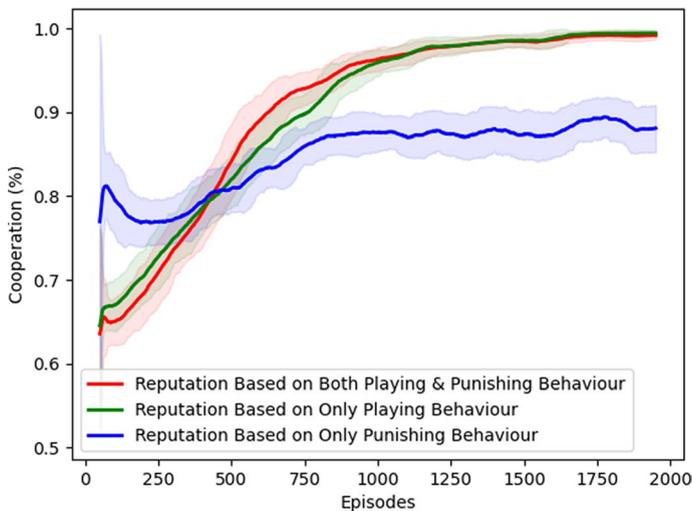
Fig. 8 Cooperation per episode achieved by calculating reputation using both playing and punishing behaviors, only playing behaviors or only punishing behaviors, within populations using third-party punishment and partner selection

However, after 500 episodes, the level of cooperation per episode achieved by populations calculating reputation using punishing behavior alone rapidly decreases compared to the levels of cooperation per episode achieved by populations considering playing behavior in agent reputations, both alone and in conjunction with punishment behavior. This suggests that only using punishment behaviors to calculate an agent's reputation is too limited to encourage a widespread emergence of cooperation within a population. This also indicates that though cooperation benefits from the presence of both playing and punishing information within agent reputations, the main value of the reputation mechanism is its ability to provide populations with a long-term view of each agents' playing behaviors.

Figure 9 illustrates that cooperation per episode is maximized when agents have access to reputational information when playing the dilemma game, but not for punishing decisions. This indicates that though access to reputational information is useful for cooperators trying to avoid exploitation by defectors when playing the Prisoner's Dilemma, it is harmful when an agent is deciding whether or not to punish another agent. This suggests that cooperation benefits from punishments based on an agent's current actions instead of their past behaviors. Though not including any reputational information in the states initially results in the highest levels of cooperation within the population, after 750 episodes it converges to a lower level of cooperation compared to when reputational information is available in the state used for playing the dilemma game. This suggests that though the inclusion of both the agents' previous actions and their reputations in the state used for playing decisions results in slower learning, as agents must learn to use a larger amount of information, it leads to higher levels of cooperation in the long term.

Including reputational information in the state used for punishing decisions or in both states results in the emergence of defection within the population. This indicates that the availability of reputational information in third-party punishment decisions is detrimental to the development of just punishment and enables defection to flourish.
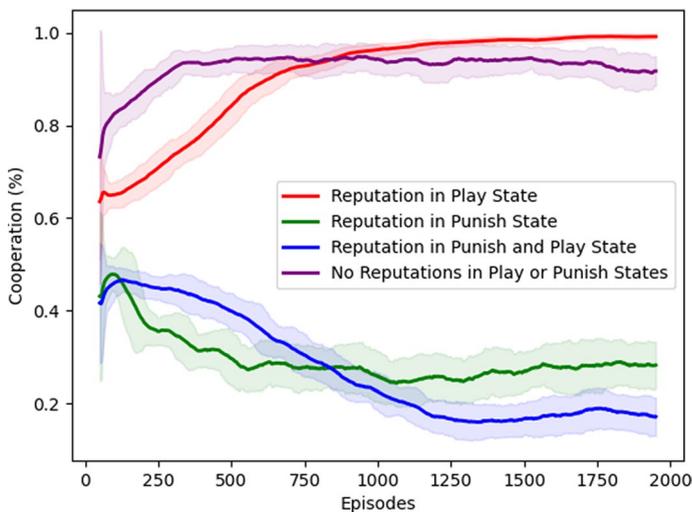
Fig. 9 Cooperation per episode achieved when reputation is included within play states, punish states and both play and punish states, in addition to the cooperation per episode achieved when reputation is not included within either the play or punish states. This is within populations using third-party punishment and partner selection

Populations using direct punishment with partner selection and reputation

Figure 10 indicates that allowing both playing and punishing behaviors to contribute to the calculation of agent reputations results in the highest levels of cooperation per episode at convergence for populations using direct punishment and partner selection and reputation. This suggests that the availability of long-term information about both the playing and punishing behaviors of each agent provides a more effective signal of agent trustworthiness compared to when reputations are calculated using an agent's playing or punishing behavior alone. However, similarly to the case of populations using third-party punishment with partner selection and reputation, the relative importance of playing and punishing information to the emergence of cooperation within populations varies during the learning process.

Up to the 500th episode, the availability of punishment behavior information is more important for the emergence of cooperation than the availability of playing behavior information. This is evidenced by the populations that determine an agent's reputation using their punishing behavior alone achieving the highest cooperation per episode within this time period. However, after 500 episodes, the cooperation per episode achieved by populations that calculate an agent's reputation using their punishing behavior alone converges to a much lower level compared to populations that calculate agent reputations based on both playing and punishing behavior or playing behavior alone. This indicates that while punishment behavior information plays an important role in the early stages of learning, its usefulness wanes in comparison to playing behavior information in the later stages of learning. This result mirrors the findings identified in the third-party punishment setting. This indicates that the relative importance of punishing and playing behavior information is similar across both types of punishment in the Prisoner's Dilemma.

Between the 500th and 750th episodes, populations calculating reputation using an agent's playing behavior alone experience a sharp, small and short-lived spike in cooperation per episode, before the levels of cooperation decrease slightly at convergence.
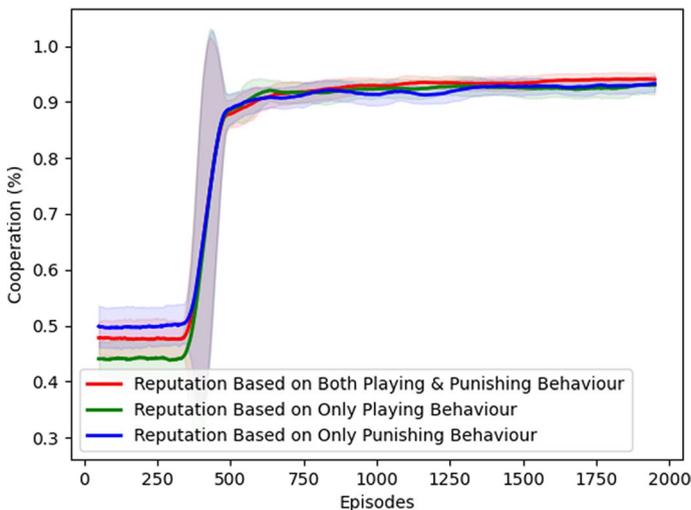
Fig. 10 Cooperation per episode achieved when reputations are calculated using both playing and punishing behaviors, playing behaviors alone and punishing behaviors alone, within populations using direct punishment and partner selection

Whereas, the cooperation per episode achieved by populations using both playing and punishing behavior information to calculate each agent's reputation continues to increase until it converges to the highest level overall. This suggests that, while information about the playing behaviors of agents plays a greater role in enabling an increase in the level of cooperation within a population in the later stages of learning, the presence of information about the punishment behaviors of an agent is still beneficial.

Unlike the third-party punishment case, Fig. 11 suggests that cooperation per episode is maximized in a population using direct punishment when reputation is not included in the definition of states used for playing or punishing other agents. Therefore, reputation has a limited ability to aid the decision making process in populations that rely on direct punishment, partner selection and reputation, beyond allowing agents to select trustworthy interaction partners during partner selection.

Similarly to the third-party punishment setting, including reputation in the state used for punishment decisions results in the emergence of defection within populations. This indicates that providing access to reputational information in the punishment step is harmful to the emergence of cooperation, regardless of the type of punishment used. Interestingly, while including reputation in the state used for the dilemma playing model does lead to some emergence of cooperation, including reputation information in both the states used by the dilemma game playing model and the punishment model results in the lowest levels of cooperation overall. This suggests that the negative impact to cooperation produced by including reputation information in the states used by the punishment model outweighs the positive affect of including the information in the states used by the dilemma playing model.
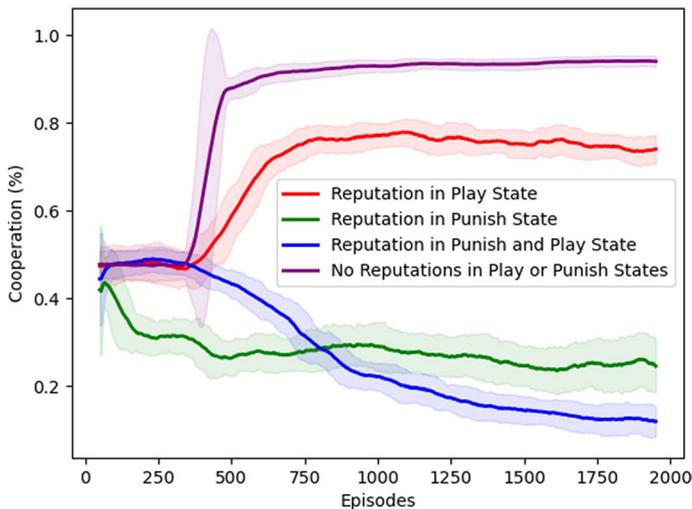
Fig. 11 Cooperation per episode achieved when reputational information is included in the state used by the dilemma playing model (play state), or when it is included in the state used by the punishment model (punish state) or when it is included in both the play and punish states. The figure also shows the cooperation per episode achieved when no reputational information is included within either the play or punish states. This is within populations using direct punishment and partner selection

## Appendix C: Determining impact of population size on cooperation

To determine the impact of population size on the experiments conducted, the experiments were also conducted on populations of size 5, 10, 15, 20, 25 and 30. As shown in Fig. 12, all the populations, regardless of size, converged to cooperation at similar rates. Figure 12a indicates that in populations where third-party punishment and partner selection are both present, larger populations achieve slightly higher levels of cooperation per episode at equilibrium compared to smaller populations. This suggests that the efficacy of third-party punishment receives some benefit from larger population sizes. However, in the case of direct punishment and partner selection, shown in Fig. 12b, there does not appear to be a meaningful pattern between levels of cooperation and different population sizes.

Though all populations converge to the same levels of cooperation in each case, Figs. 12c and 12d show that the percentage of just punishers selected at convergence is not uniform between population sizes. Figure 12c shows that smaller populations are more likely to select just third-party punishers, whereas Fig. 12d indicates that larger populations are more likely to select just direct punishers.

The result for third-party punishment may be influenced by the increased levels of defection in the 5 agent case compared to the other population sizes, which increases the
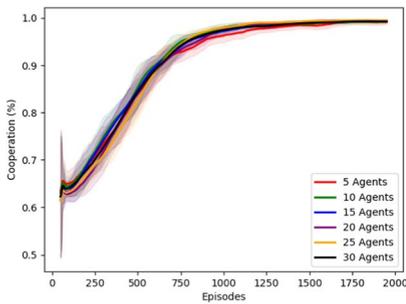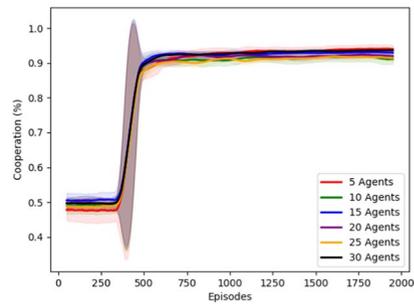
(a) Cooperation (TPP-S)   (b) Cooperation (DP-S)
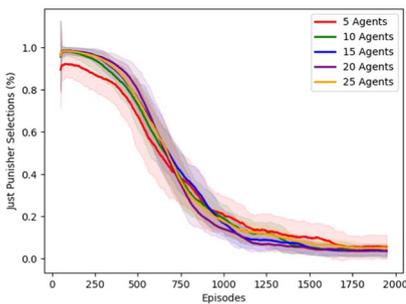
(c) Just Punisher Selections (TPP-S)   (d) Just Punisher Selections (DP-S)
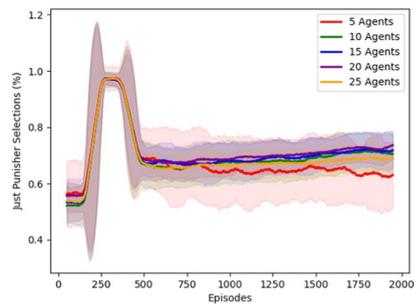
Fig. 12 Impact of different population sizes on populations using direct punishment with partner selection or third-party punishment with partner selection. The general behavior observed is similar across all population sizes

opportunity for just punishment in the latter stages of learning and therefore, increases the number of just punishers available for selection. However, this is not the case in the context of direct punishment. Figure 4b shows that all agents learn to justly punish all the time and Fig. 12b shows that approximately 10% of the population remain defectors. This indicates that larger populations do value direct punishment more than smaller populations.

Appendix D: Examining the role of conditional strategies in the emergence of cooperation

To determine whether it is sufficient for agents to act conditionally without considering punishment to promote cooperation the simulation was changed to remove the partner selection and punishment mechanisms in order to isolate the impact of conditional strategies emerging from indirect observation and direct past interactions. Figure 13 shows that in this case, cooperation fails to evolve within the population resulting in a societal collapse where all the agents defect.

This result aligns with findings from previous work studying the impact of social mechanisms on MARL populations [5, 26]. In [26], the authors show that RL agents fail to achieve cooperation in the presence of reputational information as the presence of reputational information changes the Prisoner's Dilemma into a Stag-Hunt-like game, with the RL agents converging to an inefficient equilibria in the absence of changes to intrinsic rewards or the environment. Similarly, in [5] the authors show that a population of RL agents using memory-one direct reciprocity (own and partner's previous move) fails to learn cooperation. Figure 13 further builds on these results by showing that the combination of these mechanisms still fails to achieve cooperation and results in convergence to inefficient equilibria. Together, these results indicate that reputation and direct past interactions are not sufficient for promoting cooperation. This provides further credence to the idea that reputation and direct past interactions are not individually strong factors in the
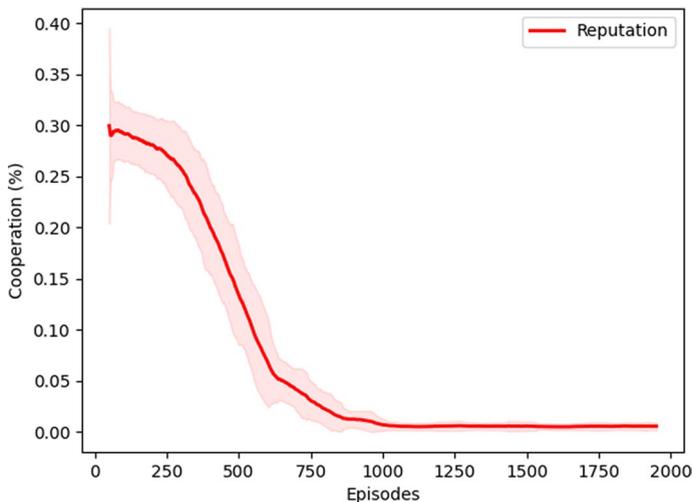
Fig. 13 Cooperation per episode for a population using past interactions and reputation fails to converge to cooperation
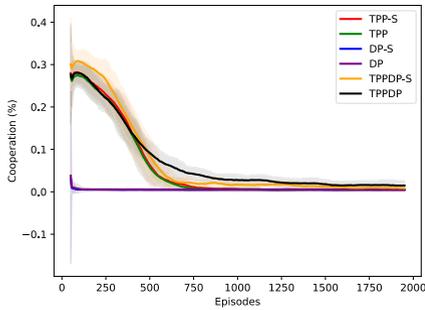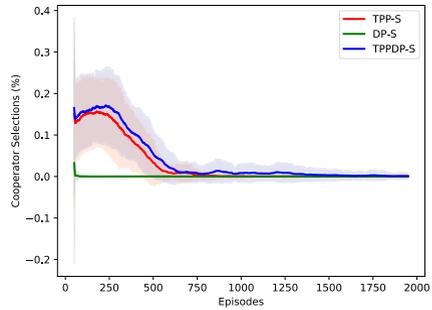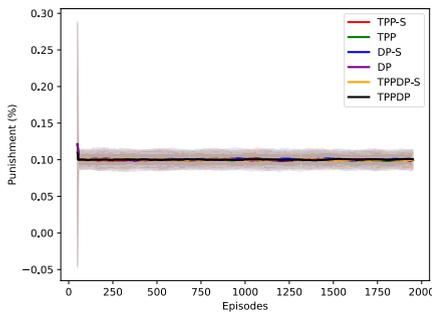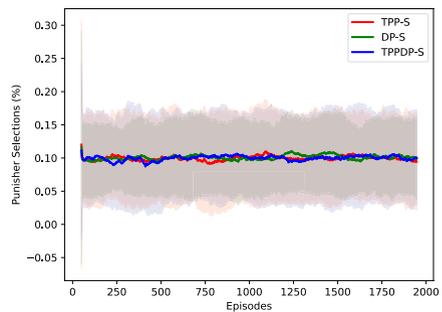
(a) Cooperation per episode   (b) Cooperator selections per episode

Fig. 14 All populations, regardless of the social mechanism combination used, fail to converge to cooperation when just punishment is costly

(a) Punishment per episode   (b) Punisher selections per episode

**Fig. 15** None of the populations learn to punish when just punishment is costly

emergence of cooperation within the population, but rather they act in tandem with other mechanisms.

Appendix E: Cooperation dynamics emerging from populations using scheme 1 to reward just punishment

The following figures show the learning dynamics emerging from populations using Scheme 1 to reward just punishment (Figs. 14, 15, 16 and 17).

Appendix F: Determining neural network size

Smaller network sizes were tested and we found that 128 nodes resulted in the most consistent and therefore, informative results. A comparison of the cooperation graphs for a network with 128 nodes compared to network with 64 nodes can be found below. At convergence, you can observe that the 95% CI for DP and DP-S is much wider for the 64 node
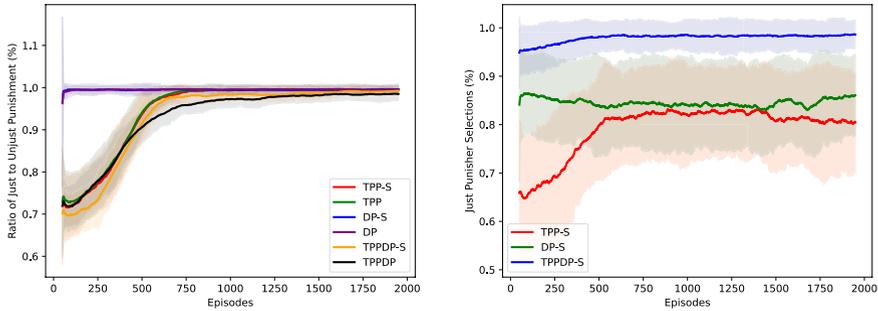
(a) Just punishment per episode   (b) Just punisher selections per episode

Fig. 16 All populations eventually always perform just punishment, but only because the population converges to defection (so the punishment is always just). This convergence to just punishment has no impact on cooperation dynamics as populations also converge to not using punishment. Populations using third-party punishment or third-party punishment with partner selection have higher levels of unjust punishment during the early stages of learning as a result of having a higher proportion of cooperators
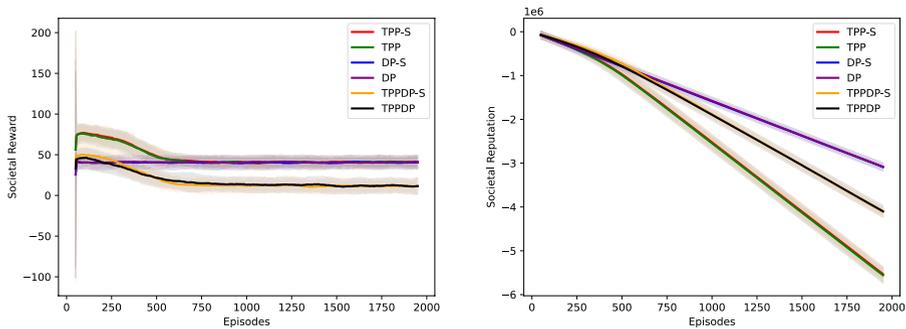
(a) Societal reward per episode   (b) Societal reputation per episode

Fig. 17 Populations using third-party punishment or combined third-party punishment initially have higher levels of societal reward compared to populations using direct punishment due to higher levels of cooperation. However, as populations converge to defection, populations using both third-party and direct punishment obtain the lowest levels of reward due to punishment occurring twice. Populations using third-party punishment or both third-party and direct punishment achieve the lowest levels of reputation as a result of greater levels of unjust punishment occurring during the early stages of learning

network compared to the 128 node network. This difference is further shown by the DP and DP-S comparison graphs below (Figs. 18 , 19).

Appendix G: Experimental code

The code used to perform the experiments used in this study can be found at the following link: https://github.com/nayanadasgupta/Understanding_Punishment_In_MAS_With_RL.

(a) 128 node neural network   (b) 64 node neural network
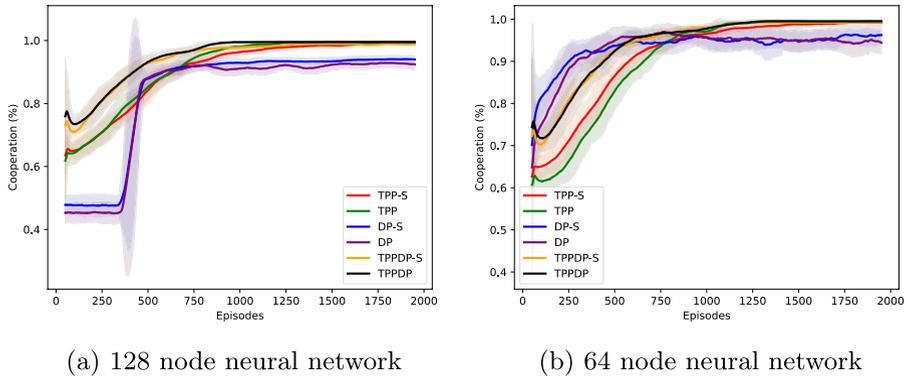
Fig. 18 Cooperation per episode. A 128 node neural network results in more consistent results

(a) Cooperation per episode (DP)   (b) Cooperation per episode (DP-S)
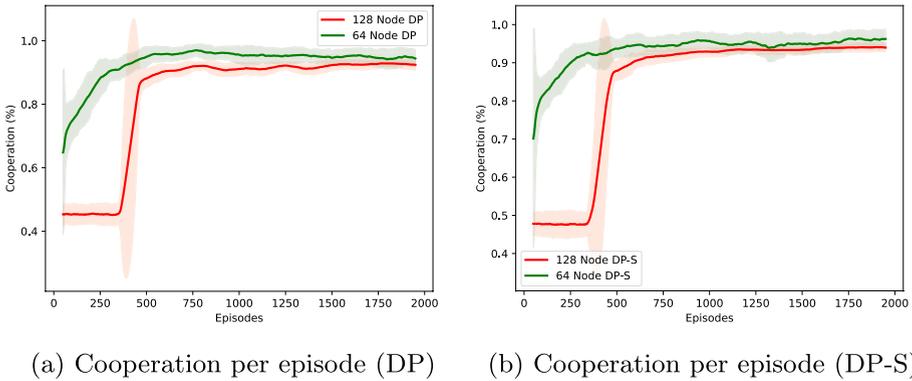
Fig. 19 Cooperation per episode for populations using direct punishment and direct punishment with partner selection. The results emerging from the 128 node network are more consistent than the results emerging from the 64 node network

Declarations

Conflict of interest The authors declare no Conflict of interest.

References

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

2. Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., & Graepel, T. (2021). Cooperative AI: Machines must learn to find common ground. *Nature, 593*(7857), 33–36. https://doi.org/10.1038/d41586-021-01170-0

3. 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

4. Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science, 211*(4489), 1390–1396. https://doi.org/10.1126/science.7466396

5. Anastassacos, N., Hailes, S., & Musolesi, M. (2020). In *Proceedings of the 34th AAAI conference on artificial intelligence (AAAI'20)*.

6. Nowak, M. A. (2006). *Evolutionary dynamics: Exploring the equations of life*. Harvard University Press.

7. Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science, 314*(5805), 1560–1563. https://doi.org/10.1126/science.1133755

8. Sigmund, K. (2010). *The calculus of selfishness*. Princeton University Press.

9. Bowles, S., & Gintis, H. (2011). *A cooperative species*. Princeton University Press.

10. Santos, F. P., Pacheco, J. M., & Santos, F. C. (2021). The complexity of human cooperation under indirect reciprocity. *Philosophical Transactions of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rstb.2020.0291

11. Santos, F. P., Santos, F. C., & Pacheco, J. M. (2018). Social norm complexity and past reputations in the evolution of cooperation. *Nature, 555*(7695), 242–245. https://doi.org/10.1038/nature25763

12. Raihani, N. (2021). *The social instinct: How cooperation shaped the world*. Random House.

13. Redhead, D., Dhaliwal, N., & Cheng, J. T. (2021). Taking charge and stepping in: Individuals who punish are rewarded with prestige and dominance. *Social and Personality Psychology Compass*. https://doi.org/10.1111/spc3.12581

14. Barclay, P., & Raihani, N. (2016). Partner choice versus punishment in human Prisoner's Dilemmas. *Evolution and Human Behavior, 37*(4), 263–271. https://doi.org/10.1016/j.evolhumbehav.2015.12.004

15. Boyd, R., & Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology, 13*(3), 171–195. https://doi.org/10.1016/0162-3095(92)90032-Y

16. Milinski, M., Semmann, D., & Krambeck, H. (2002). Donors to charity gain in both indirect reciprocity and political reputation. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 269*, 881–883.

17. Albert, M., Güth, W., Kirchler, E., & Maciejovsky, B. (2007). Are we nice(r) to nice(r) people?—An experimental analysis. *Experimental Economics, 10*(1), 53–69. https://doi.org/10.1007/s10683-006-9131-3

18. Barclay, P., & Willer, R. (2007). Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B: Biological Sciences, 274*, 749–753.

19. Sylwester, K., & Roberts, G. (2010). Cooperators benefit through reputation-based partner choice in economic games. *Biology Letters, 6*(5), 659–662. https://doi.org/10.1098/rsbl.2010.0209

20. Fu, F., Hauert, C., Nowak, M. A., & Wang, L. (2008). Reputation-based partner choice promotes cooperation in social networks. *Physical Review E*. https://doi.org/10.1103/PhysRevE.78.026117

21. Rand, D. G., Arbesman, S., & Christakis, N. A. (2011). Dynamic social networks promote cooperation in experiments with humans. *Proceedings of the National Academy of Sciences, 108*(48), 19193–19198. https://doi.org/10.1073/pnas.1108243108

22. Roberts, G., Raihani, N., Bshary, R., Manrique, H. M., Farina, A., Samu, F., & Barclay, P. (2021). The benefits of being seen to help others: Indirect reciprocity and reputation-based partner choice. *Philosophical Transactions of the Royal Society B: Biological Sciences, 376*, 20200290.

23. Takács, K., Gross, J., Testori, M., Letina, S., Kenny, A. R., Power, E. A., & Wittek, R. P. M. (2021). Networks of reliable reputations and cooperation: a review. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376.

24. Gross, J., & De Dreu, C. K. W. (2019). The rise and fall of cooperation through reputation and group polarization. *Nature Communications*. https://doi.org/10.1038/s41467-019-08727-8

25. Számadó, S., Balliet, D., Giardini, F., Power, E. A., & Takács, K. (2021). The language of cooperation: Reputation and honest signalling. *Philosophical Transactions of the Royal Society B: Biological Sciences, 376,* 20200286.
26. Anastassacos, N., García, J., Hailes, S., & Musolesi, M. (2021). In *Proceedings of the 20th international conference on autonomous agents and multiagent systems (AAMAS'21).*
27. Köster, R., Hadfield-Menell, D., Everett, R., Weidinger, L., Hadfield, G. K., & Leibo, J. Z. (2022). Spurious normativity enhances learning of compliance and enforcement behavior in artificial agents. *Proceedings of the National Academy of Sciences.* https://doi.org/10.1073/pnas.2106028118
28. Yaman, A., Leibo, J. Z., Iacca, G., & Wan Lee, S. (2023). The emergence of division of labour through decentralized social sanctioning. *Proceedings of the Royal Society B: Biological Sciences, 290,* 20231716.
29. Sen, S., Crawford, C., Dees, A., Nanda Kumar, R., & Hale, J. (2020). Effects of parity, sympathy and reciprocity in increasing social welfare. *The Knowledge Engineering Review.* https://doi.org/10.1017/S0269888920000120
30. Beckemeyer, N., Macke, W., & Sen, S. (2017). In G. Sukthankar, & J.A. Rodriguez-Aguilar (Eds.), *Autonomous Agents and Multiagent Systems.* Springer International Publishing.
31. Wu, J., Balliet, D., Peperkoorn, L. S., Romano, A., & Van Lange, P. A. M. (2020). Cooperation in groups of different sizes: The effects of punishment and reputation-based partner choice. *Frontiers in Psychology.* https://doi.org/10.3389/fpsyg.2019.02956
32. Kenward, B., & Östh, T. (2012). Enactment of third-party punishment by 4-year-olds. *Frontiers in Psychology, 3,* 373.
33. Raihani, N. J., & Bshary, R. (2015). Third-party punishers are rewarded, but third-party helpers even more so. *Evolution, 69,* 993–1003.
34. Pleasant, A., & Barclay, P. (2018). Why hate the good guy? Antisocial punishment of high cooperators is greater when people compete to be chosen. *Psychological Science, 29,* 868–876.
35. Han, T. A. (2016). Emergence of social punishment and cooperation through prior commitments. *Proceedings of the AAAI Conference on Artificial Intelligence.* https://doi.org/10.1609/aaai.v30i1.10120
36. Krellner, M., & Han, T. A. (2023). Words are not wind – how joint commitment and reputation solve social dilemmas, without repeated interactions or enforcement by third parties. *arXiv preprint* arXiv:2307.06898.
37. Beckemeyer, N., Macke, W., & Sen, S. (2017). In G. Sukthankar, & J.A. Rodriguez-Aguilar *Autonomous agents and multiagent systems.* Springer International Publishing
38. Henrich, J. (2006). Cooperation, punishment, and the evolution of human institutions. *Science, 312,* 60–61.
39. Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology, 31*(1), 169–193. https://doi.org/10.1146/annurev.ps.31.020180.001125
40. Kollock, P. (1998). Social dilemmas: The anatomy of cooperation. *Annual Review of Sociology, 24*(1), 183–214.
41. Barclay, P., & Raihani, N. (2016). Partner choice versus punishment in human prisoner's dilemmas. *Evolution and Human Behavior, 37,* 263–271.
42. Fu, F., Hauert, C., Nowak, M. A., & Wang, L. (2008). Reputation-based partner choice promotes cooperation in social networks. *Physical Review E, 78*(2), 026117.
43. Schmid, L., Chatterjee, K., Hilbe, C., & Nowak, M. A. (2021). A unified framework of direct and indirect reciprocity. *Nature Human Behaviour, 5*(10), 1292–1302.
44. Számadó, S., Balliet, D., Giardini, F., Power, E. A., & Takács, K. (2021). The language of cooperation: Reputation and honest signalling. *Philosophical Transactions of the Royal Society B: Biological Sciences, 376,* 20200286.
45. Roberts, G., Raihani, N., Bshary, R., Manrique, H. M., Farina, A., Samu, F., & Barclay, P. (2021). The benefits of being seen to help others: Indirect reciprocity and reputation-based partner choice. *Philosophical Transactions of the Royal Society B: Biological Sciences, 376,* 20200290.
46. Han, T. A., & Lenaerts, T. (2016). A synergy of costly punishment and commitment in cooperation dilemmas. *Adaptive Behavior, 24*(4), 237–248.
47. Han, T. A., Santos, F. C., Lenaerts, T., & Pereira, L. M. (2015). Synergy between intention recognition and commitments in cooperation dilemmas. *Scientific Reports, 5,* 9312.
48. Mahmoud, S., Miles, S., & Luck, M. (2016). In *Proceedings of the 2016 international conference on autonomous agents & multiagent systems* (International foundation for autonomous agents and multiagent systems, Richland, SC), AAMAS '16

49. Baumard, N., & Liénard, P. (2011). Second- or third-party punishment? When self-interest hides behind apparent functional interventions. *Proceedings of the National Academy of Sciences, 108*(39), E753–E753.

50. Molho, C., Tybur, J. M., Van Lange, P. A., & Balliet, D. (2020). Direct and indirect punishment of norm violations in daily life. *Nature Communications, 11*(1), 3432.

51. Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008). Winners don't punish. *Nature, 452*, 348–351.

52. Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics, 92*(1–2), 91–112.

53. Axelrod, R. (1986). An evolutionary approach to norms. *American political science review, 80*(4), 1095–1111.

54. Sigmund, K., Silva, H. D., Traulsen, A., & Hauert, C. (2010). Social learning promotes institutions for governing the commons. *Nature, 466*, 861–863.

55. Vinitsky, E., Köster, R., Agapiou, J. P., Duéñez-Guzmán, E. A., Vezhnevets, A. S., & Leibo, J. Z. (2021). A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings. *arXiv preprint* arXiv:2106.09012.

56. Brooks, L. C., Iba, W., & Sen, S. (2011). In *Twenty-second international joint conference on artificial intelligence*

57. Schmid, K., Belzner, L., & Linnhoff-Popien, C. (2021). In *Proceedings of the 2021 conference on artificial life (ALIFE 2021)* . 59.

58. Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature, 530*, 473–476.

59. Barclay, P., & Willer, R. (2007). Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B: Biological Sciences, 274*, 749–753.

60. Eccles, T., Hughes, E., Kramár, J., Wheelwright, S., & Leibo, J. Z. (2019). Learning reciprocity in complex sequential social dilemmas. *arXiv preprint* arXiv:1903.08082.

61. Allen, B., Lippner, G., Chen, Y. T., Fotouhi, B., Momeni, N., Yau, S. T., & Nowak, M. A. (2017). Evolutionary dynamics on any population structure. *Nature, 544*, 227–230.

62. Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science, 314*, 1560–1563.

63. Su, Q., Allen, B., & Plotkin, J. B. (2022). Evolution of cooperation with asymmetric social interactions. *Proceedings of the National Academy of Sciences, 119*(1), e2113468118.

64. Su, Q., McAvoy, A., Mori, Y., & Plotkin, J. B. (2022). Evolution of prosocial behaviours in multi-layer populations. *Nature Human Behaviour, 6*(3), 338–348.

65. Nowak, M. A., & Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature, 393*(6685), 573–577.

66. Chen, H., Zeng, Z., & Ma, J. (2020). The source of punishment matters: Third-party punishment restrains observers from selfish behaviors better than does second-party punishment by shaping norm perceptions. *PLOS ONE, 15*, e0229510.

67. Su, Q., McAvoy, A., & Plotkin, J. B. (2023). Strategy evolution on dynamic networks. *Nature Computational Science, 3*(9), 763–776.

68. Liu, Y., Zhang, J., An, B., & Sen, S. (2016). A simulation framework for measuring robustness of incentive mechanisms and its implementation in reputation systems. *Autonomous Agents and Multi-Agent Systems, 30*(4), 581–600.

69. McAvoy, A., Allen, B., & Nowak, M. A. (2020). Social goods dilemmas in heterogeneous societies. *Nature Human Behaviour, 4*(8), 819–831.

70. Sen, S., Gursel, A., & Airiau, S. (2007). In *Working notes of the adaptive and learning agents workshop at AAMAS*, vol. 7.

71. Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science, 319*, 1362–1367.

72. Rand, D. G., Armao IV, J. J., Nakamaru, M., & Ohtsuki, H. (2010). Anti-social punishment can prevent the co-evolution of punishment and cooperation. *Journal of Theoretical Biology, 265*(4), 624–632. https://doi.org/10.1016/j.jtbi.2010.06.010

73. Hadfield-Menell, D., Andrus, M., & Hadfield, G. In *Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society* (Association for Computing Machinery, New York, NY, USA, 2019), AIES '19

74. 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. (2018). Inequity aversion improves cooperation in intertemporal social dilemmas. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS'18), 31*.

75. Stewart, A. J., & Plotkin, J. B. (2012). Extortion and cooperation in the prisoner's dilemma. *Proceedings of the National Academy of Sciences, 109*(26), 10134–10135.

76. Leibo, J. Z., Zambaldi, V. F., Lanctot, M., Marecki, J., & Graepel, T. (2017). In K. Larson, M. Winikoff, S. Das, & E.H. Durfee *Proceedings of the 16th Conference on Autonomous Agents and Multi Agent Systems (AAMAS 2017)*. ACM.

77. Perolat, J., Leibo, J. Z., Zambaldi, V., Beattie, C., Tuyls, K., & Graepel, T. (2017). A multi-agent reinforcement learning model of common-pool resource appropriation. *Advances in Neural Information Processing Systems*, *30*.

78. Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning, 8*(3–4), 279–292. https://doi.org/10.1007/BF00992698