Investigating the Impact of Direct Punishment on the Emergence of Cooperation in Multi-Agent Reinforcement Learning Systems

Appendix

1 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 maximised the mean joint reward for all agents over all the repeats. The following hyper-parameter ranges were investigated during the hyper-parameter tuning process:

- Maximum Buffer Size $\in [2^x \mid x \in [11, 21)]$
- 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)
- Epsilon Decay $\in$ np.linspace(1e-4, 0.9, num=10)
- Gamma $\in [0.8, 0.9, 0.99]$
- Learning Rate $\in [0.001, 0.01, 0.1]$

	Max Buffer Size	Batch Size	Target Update	Min ϵ	Max ϵ	ϵ decay	γ	Learning Rate
Selection Model	131072	100	200	0.0001	0.8889	0.30006666666666665	0.9	0.01
Playing Model	131072	100	200	0.01	0.8889	0.30006666666666665	0.9	0.1
Punishing Model	524288	100	200	0.2	0.8889	0.5000444444444445	0.9	0.001

Table 1: Hyperparameters used for all experiments.

2 Determining Optimal Reputation and State Information Composition

Several experiments were carried out to determine what information about an agent's past behaviours should contribute to the calculation of their reputation and how this reputational information should be used, in order to maximise 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.

The first set of experiments involved comparing the levels of cooperation achieved within populations when reputation is calculated using playing behaviour alone, punishing behaviour alone

or both playing and punishing behavior. Another set of experiments evaluated the impact of allow 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.

2.1 Results

The following experimental results determine the optimal set of playing and punishing information that should be included in the calculation of agent reputations to maximise 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.

2.1.1 Populations using Third-Party Punishment with Partner Selection and Reputation

As shown in Figure 1, cooperation per episode in populations using third-party punishment with partner selection and reputation is maximised when each agent's reputation is calculated using both their playing and punishing behaviour. While calculating an agent's reputation using their playing behaviours alone results in a similar outcome to calculating their reputation using both their playing and punishing behaviours, calculating an agent's reputation using their punishment behaviours alone results in substantially lower levels of cooperation at convergence. This indicates that the presence of long-term playing information within the calculation of agent reputations is more effective at encouraging the widespread emergence of cooperation within a population, compared to the presence of long-term punishing information.

The relative importance of including playing and punishing behaviour in the calculation of reputation varies during the initial stages of learning. Between the first and the 500^{th} episode, populations calculating agent reputations using punishment behaviour 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 behaviours is a more effective signal of their trustworthiness than information about the agent's playing behaviours. Additionally, during the initial stages of learning, populations calculating an agent's reputation based their playing behaviours alone achieve a slightly more cooperation per episode compared to populations that calculate an agent's reputation based on both their playing and punishing behaviours. The delayed performance of calculating reputations using both playing and punishing behaviour 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.

However, after 500 episodes the cooperation per episode achieved by populations calculating reputation using punishing behaviour alone rapidly converges to much lower level compared to the levels of cooperation per episode achieved by populations considering playing behaviour in agent reputations, both alone and in conjunction with punishment behaviour. This suggests that only using punishment behaviours 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 2 illustrates that cooperation per episode is maximised when agents have access to reputational information within the play state, but not the punish state. This indicates that though access to reputational information is useful while avoiding exploitation when playing the Prisoner's Dilemma, it is harmful an agent is deciding whether or not to punish another agent. Though not including any reputational information in the play or punish 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 within the play state.

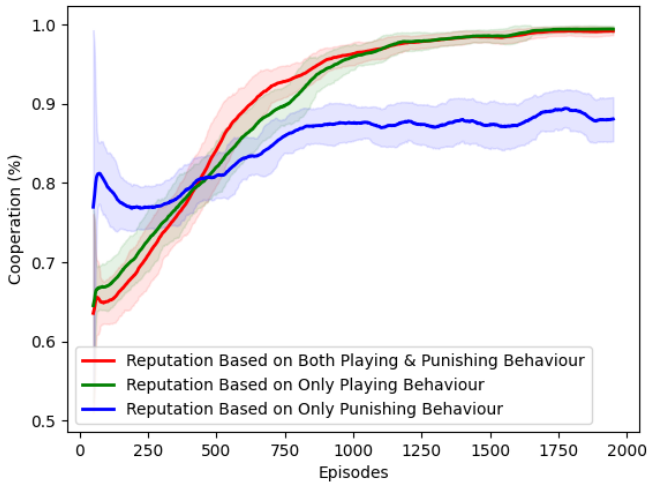

Figure 1: Cooperation per episode achieved by calculating reputation using both playing and punishing behaviours, only playing behaviours or only punishing behaviours, within populations using third-party punishment and partner selection.

This suggests that though the inclusion of both the agents' previous actions and their reputations in the play state 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 punish state or in both the punish and play 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.

2.1.2 Direct Punishment with Partner Selection and Reputation

Figure 3 indicates that allowing both playing and punishing behaviours 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 behaviours of each agent provides a more effective signal of agent trustworthiness compared to when reputations are calculated using an agent's playing or punishing behaviour 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.

Between the start of learning and the 500^{th} episode, the availability of punishing information is more important for the emergence of cooperation than the availability of playing information. This is evidenced by the populations that determine an agent's reputation using their punishing behaviour 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 behaviour alone converges to a much lower level compared to populations that calculate agent reputations based on both playing and punishing behaviour or playing behaviour alone. This indicates that while punishment information plays an important role in the early stages of learning, its usefulness wanes in comparison to playing 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 behaviour information is similar across both types of punishment in the Prisoner's Dilemma.

Between the 500^{th} and 750^{th} episodes, populations calculating reputation using an agent's playing behaviour alone experience a sharp, small and short-lived spike in cooperation per episode, before the levels of cooperation decrease slightly at convergence. Whereas, the cooperation per

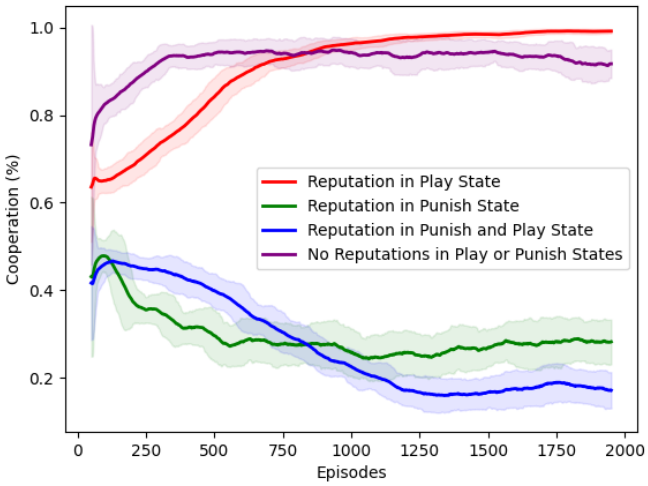

Figure 2: Cooperation per episode achieved when reputation is included within play states, punish states and both play and punish states, in addition to the cooperation per episode achieved when reputation is not included within either the play or punish states. This is within populations using third-party punishment and partner selection.

episode achieved by populations using both playing and punishing behaviour 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 behaviours of agents plays a greater role in enabling growth of cooperation within a population in the later stages of learning, the presence of information about the punishment behaviours of an agent is still beneficial.

Unlike the third-party punishment case, Figure 4 suggests that cooperation per episode is maximised in a population using direct punishment when reputation is not included in either the playing or punishing states. Therefore, reputation has a limited ability to aid the decision making process in populations using 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 punish state 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 play state does lead to some emergence of cooperation, including reputation in both the play and punish states results in the lowest levels of cooperation overall.

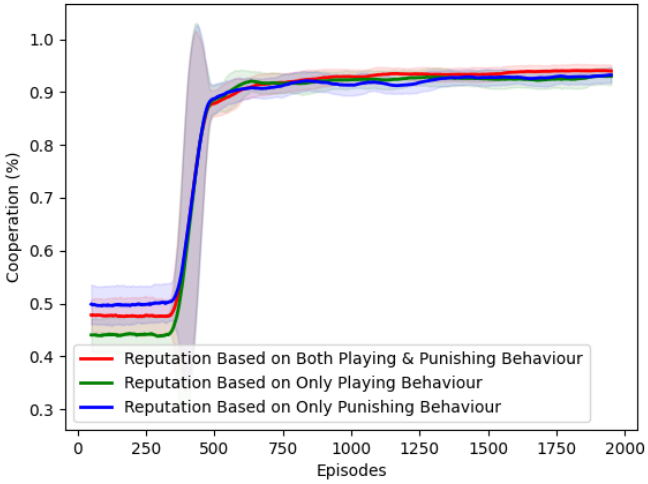

Figure 3: Cooperation per episode achieved when reputations are calculated using both playing and punishing behaviours, playing behaviours alone and punishing behaviours alone, within populations using direct punishment and partner selection.

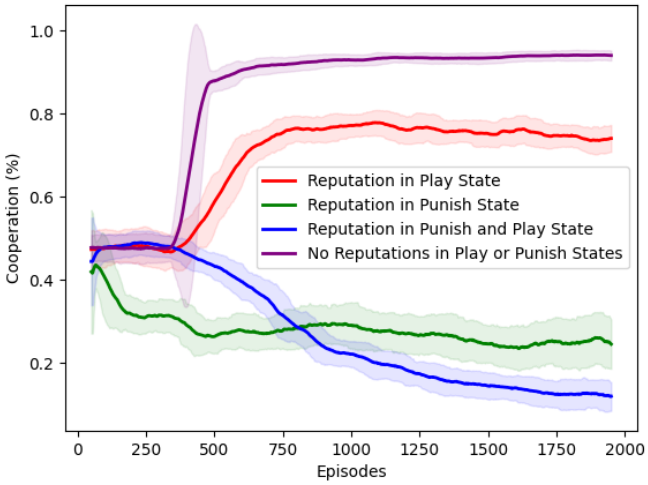

Figure 4: Cooperation per episode achieved when reputational information is included in the play states, punish states, both the play and punish states and when no reputational information is included within either the play or punish states. This is within populations using direct punishment and partner selection.