Augmenting Reinforcement Learning to Enhance Cooperation in the Iterated Prisoner's Dilemma

Grace Feehan[a] and Shaheen Fatima[b]
Loughborough University, Epinal Way, Loughborough, U.K.

Keywords: Iterated Prisoner's Dilemma, Interaction Structure, Multiagent Reinforcement Learning, Mood, Cooperation Index.

Abstract: Reinforcement learning algorithms applied to social dilemmas sometimes struggle with converging to mutual cooperation against like-minded partners, particularly when utilising greedy behavioural selection methods. Recent research has demonstrated how affective cognitive mechanisms, such as mood and emotion, might facilitate increased rates of mutual cooperation when integrated with these algorithms. This research has, thus far, primarily utilised mobile multi-agent frameworks to demonstrate this relationship - where they have also identified interaction structure as a key determinant of the emergence of cooperation. Here, we use a deterministic, static interaction structure to provide deeper insight into how a particular moody reinforcement learner might encourage the evolution of cooperation in the Iterated Prisoner's Dilemma. In a novel grid environment, we both replicated original test parameters and then varied the distribution of agents and the payoff matrix. We found that behavioural trends from past research were present (with suppressed magnitude), and that the proportion of mutual cooperations was heightened when both the influence of mood and the cooperation index of the payoff matrix chosen increased. Changing the proportion of moody agents in the environment only increased mutual cooperations by virtue of introducing cooperative agents to each other.

1 INTRODUCTION

Models of human social behaviours, and the socio-cognitive mechanisms that underlie them, have become a fast focus of researchers in the field of computer science in recent years. In particular, they are of interest both for the improvement of existing algorithms (such as for the focusing of attention in visual search; (Belkaid et al., 2017) but also in investigating individual and group behaviours (social network simulacra being one example). One testing ground for the development of such artificial mechanisms are social dilemmas, as they are relatively well understood, simple and easy to constrain or build on. In these we find strong methods of studying interpersonal behaviour.

This work builds on previous research on how typically reward-focused reinforcement learning (RL) algorithms, utilised by a network of agents engaged in the Iterated Prisoner's Dilemma (IPD), can be altered using computational socio-cognitive models to encourage increased cooperation. One of the newer of these in recent literature has been a model of mood, with research suggesting that a key determinant of the emergence of cooperation within moody dilemma playing is the interaction environment (or network structure; Collenette et.al., 2017a). The model has so far been tested primarily in mobile networks of moving agents; we instead seek to more clearly establish how a static and more standardised structure of interaction affects the moody model's influence on cooperation. **The main objective** of this paper is to find methods of increasing cooperation in RL agents playing the IPD. To do this, we evaluate the behaviour of moody learning agents in a static, regular, multiagent environment whilst playing against more traditional RL agents - data on which has not been explicitly presented in past research. We first summarise the published research on a human-based model of mood and then examine how behaviour observed in the novel environment compares. We hypothesised that results may follow the data in Collenette et.al. (2017a), and demonstrate that increased interaction rates dampen cooperation gains, but still show an increase in cooperation over unaltered RL. We also investigate **two additional factors** for increasing cooperation rates; the proportion of opponents in the environment and a measure of how cooperative the reward schema is. **Results of our simulations demonstrated** that cooperation can indeed

[a] https://orcid.org/0000-0002-8629-1800
[b] https://orcid.org/0000-0002-6068-2942

Feehan, G. and Fatima, S.
Augmenting Reinforcement Learning to Enhance Cooperation in the Iterated Prisoner's Dilemma.
DOI: 10.5220/0010787500003116
In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022) - Volume 3*, pages 146-157
ISBN: 978-989-758-547-0; ISSN: 2184-433X

be increased by the latter of these factors by up to 15%, whereas the former does not. Our experimental data are discussed in the context of structures of interaction, the variables the model utilises, and possible changes the model's implementation may need in this environmental setting. **The primary contributions** of this research are that we provide direct evaluation of the moody strategy in static networks (previously undetailed), test the moody model against varying quantities of traditional RL opponents to evaluate behavioural transference, and test varying payoff matrices which facilitate comparisons with external human participant data. The **Python code repository** for this paper is available on request.

Beginning with *Section 2*, we outline the computational problem space covered by the IPD, and the justification for using it in uncovering social agent behaviour. We also describe the RL algorithm *SARSA*, and introduce the notion of interaction structures and how this relates to the moody model to come. Then, in *Section 3* we describe this model of mood (*mSARSA*) as it is primarily outlined in Collenette et.al. (2017b), with deeper presentation of its relevant past performance. We also provide the motivation, based on this literature, for our own experiments. *Section 4* provides the detail of the novel testing environment, parameters and hypotheses. Finally, *Section 5* gives the data from these experiments, and *Section 6* gives the subsequent analysis, comparisons and conclusions. Thereafter we propose suggestions for the improvement of the model.

2 BACKGROUND

2.1 Prisoner's Dilemma

The Prisoner's Dilemma is a social dilemma in which players have two behavioural choices available to them, and typically must choose without direct communication. Cooperate with the partner, and potentially be exploited, or defect (play selfishly) against the partner, which (if mutual) will lead to a lower round payout than with mutual cooperation. The Nash Equilibrium of the dilemma, due to the payoff structure (*Table 1*), is to defect - but if both players opt to do this, their score is diminished in comparison with what could be achieved. In addition, when the game is played over many rounds (as in the IPD), several issues arise. The strong incentive to defect leads networks to obtain substantially lower payoffs as a whole than sustained mutual cooperations would provide. There is also a range of problems involving how to encourage partners to both trust and *continue* playing the game with you when you repeatedly defect (Wilson and Wu, 2017). This lends itself to the motivation for finding mechanisms that encourage the evolution of cooperation in artificial agents.

Table 1: Payoff matrix for the two-player Prisoner's Dilemma game, as used in Collenette et.al. (2017b) and throughout classic literature. The cooperation index of this matrix (see *Section 3.3*) is 0.4.

		Partner B (Right)	
		Cooperate	*Defect*
Partner A (Left)	*Cooperate*	**3,3**	**0,5**
	Defect	**5,0**	**1,1**

The payoff values presented in *Table 1* have been used extensively throughout IPD research under different names; **Temptation** (or **T**) for a DC outcome, **Reward** (or **R**) for a CC outcome, **Punishment** (or **P**) for a DD outcome and **Sucker** (or **S**) for a CD outcome. In our work, we also refer to these as the *Exploiter*, *Mutual C*, *Mutual D* and *Exploited* payoffs respectively as they provide simpler intuition.

Overall, we find a suitable research environment given two reasons. Firstly, there is well-established precedent that human networks playing the game do not defect in the same manner as many game-playing strategies (Barreda-Tarrazona et al., 2017), and are more heterogeneous in their behaviours (Fudenberg et al., 2012). Secondly, there is large potential for artificial players to improve from payoff-incentivised defection with the introductions of novel additions (as they can already display greedy behaviour; Collenette et.al., 2017b). RL algorithms such as *SARSA* have been used to learn behavioural policies for the IPD (Collenette et.al., 2017b; Yu et.al., 2015; (Gao, 2012), and in general are of interest to computational neuroscience regarding human behaviour (Shteingart and Loewenstein, 2014); making them a suitable methodology for investigating humanlike cooperative behaviour. In the following section, we detail *SARSA* as it relates to social agent research, and then in *Section 4* we outline how it has been extended further with a moody model, noting additional factors that may influence cooperation behaviour.

2.2 Interaction Structure

One of the primary sources demonstrating the effect a model of mood has on the emergence of cooperation within social dilemmas is Collenette et.al. (2018), a study which concludes that the structure of network connections and interactions influences how behaviours emerge in said network. Here we will outline some of the terminology involved and how each relates to the research at hand. Interaction structures can be defined as the way agents interact with each other; in the context of the IPD, interactions occur between two players when they play a round of the dilemma game. In our experiments and in much of the research summarised in *Section 3*, this is extended to involve multiple pairs of players in a group, or a network. Networks can be given in the form of a graph, where agents in these networks are represented as nodes and interaction relationships between agents are represented as edges between nodes. There are two main forms of network related to the current research (mobile and static), and there are many structural variations on the graph which represents how well connected agents are.

In the mobile environments used in the majority of the model literature, agents move freely around arenas with obstacles (dependent on the network type), intending to replicate the dynamic and uneven interactions of more natural societal networks. These environments have equivalent static networks with differing topographical structures; in Collenette et.al. (2018), for example, the authors include *small world* (characterised by high clustering), *fully connected* (where all nodes are connected to all other nodes), *random* (as the title suggests) and *regular* networks (where each node has the same number of edges). The equivalent in our own experiments would be a static regular network. As the literature discussed here describes the effect of increasing mobility, and there is some disagreement amongst the papers on the effect of network structure, evaluating both *SARSA* and *mSARSA* in a static regular network may add more clarity to structural effects. In the learning algorithm *SARSA*, agents update world knowledge through having game interactions. This makes the nature of those interactions, how many interactions agents have in the course of a learning episode, and what information those interactions provide, critical to this learning.

2.3 *SARSA*

State-action-reward-state-action (*SARSA*) is an on-policy RL algorithm - one that is used throughout the literature pertaining to the mood model discussed here. States, in the case of default *SARSA* as it is used in this paper, are the histories of interactions with each partner - the length of which depending on the memory size afforded to the agent. In our implementation, up to seven items can be remembered or for larger items (like the memory of what *both* players did), 3 or 4 ‘chunks’. Actions are the behaviours available (i.e. *cooperate* or *defect*), and rewards are given by the payoff matrix used, depending on the outcome of each individual interaction. Both the default *SARSA* and the *mSARSA* variant use the ε-greedy behaviour selection mechanism. Algorithmic details and pseudocode for *SARSA* can be found in (Sutton and Barto, 2018). The equation for updating Q (the learned value) for each state-action combination is given below as it is pertinent for comparison with the mood-augmented version presented in *Section 3.1*.

Let s represent the state, t denote the current timestep of a learning epoch, a the action taken in that state, where α is the learning rate (typically 0.1), γ is the discount factor (typically 0.95), and r is the reward received:

$$\begin{aligned} Q(s_t,a_t) = Q(s_t,a_t) + \\ \alpha[r_{t+1} + \gamma Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)] \end{aligned} \tag{1}$$

Originally selected due to its on-policy nature conforming with the on-policy characterisation of mood (Collenette et.al., 2017b), it has the added benefits of being well-established, easy to implement and computationally light when the state space is smaller (Collenette et.al.’s methodology yields approximately 216 states, for instance). Application of reinforcement learners is of interest to those researching artificial social and cognitive mechanisms in humans; Shteingart et.al. (2014) discusses the benefits of model-free temporal difference learners as practical methodologies for examining and recreating human-like behaviour, but also the potential lack of intricacy and depth that model-free algorithms have when attempting to recreate behaviour altered by chemical neuromodulation (perhaps particularly of relevance to mood and emotion modelling).

Still, *Q-learners* are being pursued as viable methods to model different variations of human processing (Lin et al., 2019)) and continue to be of use to interdisciplinary research. Bearing these limitations in mind, we can consider how augmentations to *SARSA* may also begin to pursue more complex humaniform modelling. One particular field of interest in regards to this is mood, where there exists a large body of literature on how mood influences dilemma reasoning (Palfai and Salovey, 1993) and cooperative behaviour (Proto et al., 2017); (Hertel et al., 2000). We will cover one particular computational model of mood that has been integrated into the *SARSA* learning algorithm.

3 RELATED LITERATURE

3.1 Mood Model and Integration with *SARSA*

Mood is reliably, if reductively, defined as a task- and partner-invariant affect spectrum which influences other cognitive processes, dissociated more from current events with longer-term effects (Collenette et.al., 2017b). It is a closely interdependent subsystem with emotions[1] and biases many aspects of human perception.

The moody *mSARSA* model under analysis here is primarily outlined in Collenette et.al. (2017b), and employs a central mechanism by which mood is a real number between 1 and 100. Higher mood values are coded as more risky and cooperative, with lower moods as more rational and defective; in Collenette et.al., extreme moods (above 90 and below 10) are implied as characterising mood disorders (mania and depression, respectively). Mood updates are made solely around how an agent perceives its current payoff (which is adjusted using the *Homo Egualis* model, comparing the reward with the opponent's reward; (Fehr and Schmidt, 1999), relative to the average payoff previously attained. If an agent perceives itself to be doing poorly in comparison with its past, its mood will decrease. In Collenette et.al. (2018), this process is slightly adjusted to facilitate higher moods decreasing more readily (in line with the ease of fluctuation at lower moods). The mood-altered equation for updating *Q* for each state-action pair is provided below (adapted from Collenette et.al., 2017b).

$$Q(s_t,a_t) = Q(s_t,a_t) + \alpha[r_{t+1} + \gamma\Psi - Q(s_t,a_t)] \quad (2)$$

The rule for the estimation of future rewards (Ψ) is encapsulated in *Equations 3* through *5* (reproduced from Collenette et.al., 2017b), where m_i^t is the mood of a given agent i at timestep t (a $\mathbb{R}$ between 0 and 100). Mem_i^a is the vector of the set of rewards previously obtained by that agent when using action a, and $|Mem_i^a|$ is at maximum 20. $Mem_i^a(0)$ returns the most recent reward.

$$\alpha_i^t = (100 - m_i^t)/100 \quad (3)$$

$$\beta_i^t = ceil(|Mem_i^a(n)|/\alpha_i^t) \quad (4)$$

$$\Psi = (n \sum_{0}^{\beta} Mem_i^a(n))/\beta_i^t \quad (5)$$

Mood constrains the depth (denoted with n above) at which memory is consulted for the average past reward, but also controls the value ε in ε-greedy exploration. When an agent's mood is below 30 and they cooperate, or above 70 and they defect, ε increases to 0.9 for that turn and a move is re-selected. Mood is updated and maintained by the series of *Equations 6* through *8* (reproduced from Collenette et.al., 2017b). Let t denote the current timestep of a learning epoch, p_i^t return the payoff of agent i in that timestep, μ_i^t denote their average payoff over the elements in Mem_i^a, and m_i^t denote their mood. Let j denote agent i's opponent, and let $\alpha = \beta$ (as in Collenette et.al., 2017b):

$$\alpha_i^t = (100 - m_i^{t-1})/100 \quad (6)$$

$$\begin{aligned}\Omega_{i,j}{}^t = \mu_i{}^t - \alpha_i{}^t \cdot max(\mu_j{}^t - \mu_i{}^t, 0) - \\ \beta_i{}^t \cdot max(\mu_i{}^t - \mu_j{}^t, 0)\end{aligned} \quad (7)$$

$$m_i{}^t = m_i{}^{t-1} + (p_i{}^t - \mu_i{}^{t-1}) + \Omega_{i,j}{}^{t-1} \quad (8)$$

Algorithm 1: *mSARSA* Pseudocode (Adapted from Collenette et.al., 2017b).

```
initialise all Q(states, actions) arbitrarily;
for each episode do
    initialise all states;
    Choose an action using the policy derived from Q ($\varepsilon$-Greedy);
    for each episode step do
        Take the chosen action, observe the reward and the new state reached;
        if mood $\geq$ 70 and action = 'D' OR mood $\leq$ 30 and action = 'C' then
            re-select an action under a higher $\varepsilon$ value (0.9)
        Choose next action using the policy derived from Q ($\varepsilon$-Greedy);
        Estimate future reward using Equations 3 through 5;
        Update Q(s, a) using Equation 2;
        Update mood using Equations 6 to 8 ;
Until terminal step;
```

As *Algorithm 1* demonstrates, moody alterations to the standard *SARSA* framework are mostly additive, with the exception of one variable replacement in the *Q*-update equation. At each stage of learning mood value is considered, creating an all-permeating integration to decision-making. The purpose of the system design is a thoughtful combination of a few motivations: it models evidence-based mood in a learner, where previous focus had been primarily on models of emotion; it intends to expand understanding on how human-like mood (and emotion, where applicable) informs decisions in social dilemmas; and it appears to facilitate increased rates of cooperation in certain situations, as we will evidence from past research.

One final important aspect of the algorithm to note is that *Equation 8* provides the first iteration of the mood update function (MU1) that is utilised in Collenette et.al. (2017b), which this paper will make a direct comparison with in terms of data. There is a second iteration of this equation (MU2, see *Equation*

[1] Emotions, which are more temporally distinct, varied, and *can be* task-relevant (Zulfiqar and Islam, 2017) also have an influence on decision-making; despite their relevance, however, emotions are extraneous to the scope of the current paper.

9), which was developed and deployed in Collenette et.al. (2018) to facilitate faster mood changes when poor outcomes are received at higher mood levels.

$$m_i^{\,t} = m_i^{\,t-1} + (p_i^{\,t-1} - \Omega_{i,j}^{\,t-1}) \tag{9}$$

It is important to note that our implementation (see *Section 4*) uses the *former* of these equations (*Equation 8*) in its calculations, for two main reasons. First, it means we utilise the exact structure of the version used in Collenette et.al., (2017b), which is the only source of precise cooperation data for the algorithm specifically (and not a mixed population as a whole, as presented in Collenette et.al., 2018). Secondly, initial testing with MU2 indicated that it took a lot longer for the algorithm to run to conversion - if at all, as it potentially exhibited cyclic behavioural loops. It did, however, manipulate the mood value much more effectively than MU1 appeared to do, with the behaviour of both possibly exaggerated by the nature of our environment used. Though full details on our observations cannot be evaluated here due to constraints, this may be examined in further work - the limitations and critique of MU1 where it is relevant to the current experiments are detailed clearly in *Section 6*.

3.2 Moody versus Non-Moody

Firstly, it is quite apparent that the addition of the mood model in previous research scenarios is a positive one; networks with moody agents achieve greater proportions of mutual cooperation at behavioural convergence than their non-moody (or even emotional) counterparts. Moody agents do not necessarily always attain the greatest utility[2] but the mechanism does seem to reduce the self-interest of otherwise greedy *SARSA* agents. These agents are typically rendered using Stage (Vaughan, 2008) and perform random walks, usually in an environment with central obstacles. On encountering an opponent, they perform a round of the Prisoner's Dilemma (or other game, as relevant) and then move on. It is worth bearing in mind this setup throughout the following analysis, as the authors repeatedly comment on how reduced rates of uncertainty in repeated interaction introduces difficulty in sustaining cooperation and removes the threat of punishment for defecting.

Initial work combining the mood system with the Ortony, Clore and Collins (OCC) Model of Emotions (Ortony et al., 1988), (Clore and Ortony, 2013) demonstrated a high rate of effectiveness in the IPD. In Collenette et.al. (2017b), a three-emotion framework saw an increase in the peak proportion of mutually cooperative outcomes from approximately 35% to 100% in some scenarios when the mood model was added. With the scenarios representing different proportions of starting moods (between low, neutral and high) in the environment, it is clear that in initial testing high mood networks happily cooperate with one another. As discussed later in the paper, however, these are not resilient outcomes, and are vulnerable to exploitative crashes against pure defectors (versus low mood groups). In these resilience tests, the percentage of mutual cooperation rose at best by approximately 16%, which is still a marked improvement over the 5% initial rate. This paper also demonstrates the positive visual correlation between mood value and percentage of mutually cooperative outcomes, with both rising steadily as more interactions occur.

In a highly mobile environment, the model facilitated an increase in the proportion of mutually cooperative outcomes by at least 40% over the *SARSA* equivalent; both when opponents were recognisable and when their mood labels were directly observable (Collenette et.al., 2017b). This increase was even greater when agents possessed no information about their opponents, with regular *SARSA* agents attaining only 1.7% mutually cooperative outcomes and their moody counterparts attaining (at greatest) 78%. There were equivalent decreases in the proportion of mutually defective outcomes in these conditions where *mSARSA* performed well.

In more explorative network research, the model was included amongst a very wide variety of other strategies in a broader selection of topologies. In one paper, though there is no per-strategy breakdown provided in the data regarding cooperation or utility gathered, mutual cooperation was maintained at approximately 49% in a mixed strategy environment (Collenette et al., 2018b). Importantly, this paper draws several key conclusions about the effect network connectivity (and the equivalent physical environments) has on both cooperation and dilemma playing. Networks with random connectivity facilitate the greatest proportion of mutual cooperation and average payoffs to agents, with the mobile environment equivalent providing the same in terms of cooperation, but performing only approximately as well as the empty and small world environments. Though the effect on the mood model specifically is unknown due to lack of strategy-by-strategy breakdown, it is important to note that our environment is most comparable with the regular static network in Collenette et.al. (2018) (the mobile equivalent of which is utilised in the ma-

[2] In Collenette et.al. (2018), *Tit-for-Tat* was identified as the most successful strategy despite differing environments and network structures.

jority of the tested environments across the papers discussed). In the terms of connectedness, the regular static network performed second worst out of the four constructions (demonstrated lower cooperation and lower average payoff attained) but attained higher scores on both metrics in comparison to its mobile equivalent.

One final paper (Collenette et al., 2017a) combines two of these methodologies - the emotional-moody non-learning model and a variety of different network types, similar to Collenette et.al. (2018). Again, mood shows a strong positive influence on cooperation levels over just the OCC framework alone. Emotional agent networks sustained cooperation almost regardless of length of interaction, with starting level of cooperation having a much larger effect than the distribution of admiration thresholds (one of the three emotions used) on the level of cooperation in the system. Most of the results where initial cooperation was set to 50% resulted in mutual cooperation proportions within the range of 25% to 40%. On the other hand, the inclusion of mood increases this range to between approximately 82% to 100%. Resilience results are similar to those seen in Collenette et.al. (2017b), demonstrating low moods as more resistant to pure defectors in the long run.

3.3 Cooperation Index

Throughout the moody algorithm literature, variations on the traditional payoff matrix (*Table 1*) have not been tested. This is an important factor for an algorithm modelling humanlike traits, as there exists work that demonstrates how relational differences between the payoffs influences human cooperation rates in the IPD. In one paper (Wrightsman et al., 1972), the authors summarise and discuss initial research into a *cooperation index* (henceforth referred to as *K*, as in Colman et.al., 2018), or formula for characterising the intensity of the conflict the structure of the payoff matrix creates. This text is rather old, but summarises past work regarding human interactivity with *K* well; namely, cooperative behaviour appears to increase as *K* increases, with less clear-cut conclusions made regarding its influence on defections (something the authors suggest is primarily influenced by the **T** payoff alone). The equation for *K*, where each input has been defined previously in *Section 2.1*, is given in *Equation 10*.

$$K = (\mathbf{R} - \mathbf{P})/(\mathbf{T} - \mathbf{S}) \tag{10}$$

Other research continues to use *K* to evaluate payoff matrices with humans and compare it to other models. *K* has been shown to be a strong cooperation predictor when values of *K* are distributed from 0.1 to 0.9 (Hristova and Grinberg, 2005). These authors again found a consistent increase of cooperation as *K* increased, and an influence of contextual *K* (that is, playing other IPD games with high *K* values around a specified target game with a differing *K*) on cooperation levels. The authors also create a computational model that utilises subjective expected utility theory, combined with weightings for the importance of each outcome type (calculated via average of past payoffs) to replicate the trends shown in the human data, with success. One paper reiterates the human conclusions from this previous work in one of the most recent references to *K*, providing solid basis of connection between *K* and cooperation (Colman et al., 2018).

3.4 Summary and Literature Gap

The evidence discussed thus far indicates that the model has a strong propensity for inducing mutual cooperation between homogeneous agents, with indirect data suggesting it possibly assists in the formation of a stable level of cooperation in heterogeneous strategical groups. It has been flexibly combined with RL and frameworks of emotional behaviour alike, and in both cases has allowed for the increase in cooperation. Highly connected networks reduce the effectiveness of the model as opposed to randomly connected networks, and with mobility added on top of this, cooperation is even further diminished in mixed groups. There is also evidence to suggest that in homogeneous groups, the influence of network orientation is much less important than variables such as starting mood (though this is in non-learning populations).

Overall, there is a lack of specific, isolated data on the performance of the moody learner in a static, homogeneous environment with regular connectivity, and no data whatsoever on how moody agents behave in environments with varying distributions of opponents or different payoff matrices. The experimentation in *Section 4* seeks to partially rectify this. For these reasons, we will test *mSARSA* against itself, replicating the parameters thoroughly tested in Collenette et.al., (2017b) but under a more consistent interaction structure. We will also substantially evaluate the two contributing factors outlined (opponent density and *K* value) in order to provide a thorough examination of behavioural outcomes. If we can establish a baseline of whether *mSARSA* behaves similarly and reliably in our environment in comparison with mobile counterparts, we then not only lend weight to the model's robustness, but also set precedent for further experimentation.

4 EXPERIMENTAL EVALUATION

4.1 Design

The environment used for all three of the following experiments is a regular 2D grid network of 25 agents who do not change positions between rounds. Grid positions begin from the bottom left (pos. 1) and increase vertically northwards, filling out each column before moving to the next (i.e. top left is pos. 5, top right is pos. 25). In a single round of the IPD, agents interact with as many available partners in the four *von Neumann* neighbourhood positions, *north, south, east* and *west* (clockwise, in this order). In ***Interaction Structure*** experiments, all agents in the grid are of the same strategy (mSARSA or SARSA). In ***Strategy Proportion***, most agents in the grid are initially SARSA, with the proportion of agents increased gradually as documented in *Table 2*. In ***Payoff Schema***, there are 12 mSARSA agents in the environment to 13 SARSA agents, using the *F* agent arrangement from the previous experiment (*Table 2*).

This structure was designed in MESA (Team, 2021), recreating grid environments from other spatial social dilemmas (Yu et al., 2015). **Data presented are the results of only the inner square grid of agents (excluding all border players)** - this provides more standardised analysis as these agents will have the maximum of four partners. Each experiment consists of 5 learning episodes (experimental periods in which internal variables are maintained), each consisting of 10,000 time steps. This value was chosen to compromise between efficient runtime and allowing for any late-phase change of behaviour. Data presented averages over these learning episodes.

4.2 Parameters

In the base experiment, *mSARSA* will be tested (as in Collenette et.al., 2017b) using three levels of state information - the opponent's last move, the opponent's ID, and the opponent's mood label - each cumulative with the last, under the names *Stateless*, *Agentstate* and *Moodstate*. The parameter *mA* is intended to reflect the 'extent to which mood influences behaviour' (Collenette et.al., 2017b); this takes the form of altering moody behavioural thresholds chosen each round and how epsilon is changed. We will test this parameter with values 0 through 0.8, in increments of 0.4. In the *Strategy Proportion* and *Payoff Schema* experiments, *mSARSA* utilises just the *Moodstate* and *mA*=0.8 parameters.

In the basic test of *Interaction Structure*, agents solely play against opponents of their own strategy in the grid. In the two experiments following this, *mSARSA* agents play the IPD with *SARSA* agents. The positions and proportions of agent types in the environment for *Strategy Proportion* tests are outlined in *Table 2*, with the *Payoff Schema* experiments utilising the **Condition F** structure from this same table.

Table 2: Parameters used for the *Strategy Proportion* experiment. The first row notes the grid locations of *mSARSA* agents in each condition, and the second row displays what percentage of the whole population the *mSARSA* agents occupy.

Condition	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Grid Locations	1	9	7, 9, 17, 19	1, 5, 21, 25	C, D	E, 6, 10, 16, 20	F, 2, 4, 22, 24	G, 3, 11, 15, 23
Network Proportion (%)	4	4	16	16	32	48	64	80

The independent variable changed in *Payoff Schema* is the payoff matrix used, varied both in actual value size and in size of cooperation index (*K*) - a measure used in both Wrightsman et.al. (1972) and Colman et.al. (2018) (see *Table 3*). These can be broadly categorised into three sections; small-value payoffs (between 0 and 1) with small, medium and large values of *K*, the traditional IPD payoff scheme (**Condition 4**) and then large-value payoffs (between 0 and 100), again with small, medium and large values of *K*. This should provide some insight not only into influence of increasing *K*, but also the actual value of payoffs themselves. We also investigate the influence of increasing *K* when the T and R payoffs remain the same - the payoffs used in this portion of the experiment can be found in *Table 4*. *SARSA* utilises $\alpha = 0.1$, $\gamma = 0.95$ and a linear decay mechanism for ε that starts at 0.99 and decays linearly to 0.1 mid-way through each episode. It also uses a state information profile closest to *Stateless* with a single-step memory.

Table 3: Parameters used for the first section of the *Payoff Schema* experiment. Each column provides the payoff matrix utilitsed by each numbered condition. *K* is the value of the Cooperation Index for that payoff matrix column.

Outcome	**Payoff Value Used**						
Condition	**1**	**2**	**3**	**4**	**5**	**6**	**7**
T	0.8	1	1	5	99	56	99
R	0.6	0.9	0.9	3	52	37	98
P	0.5	0.1	0.1	1	51	16	1
S	0	0.6	0	0	0	14	0
K	*0.12*	*0.50*	*0.80*	*0.40*	*0.10*	*0.50*	*0.97*

4.3 Hypotheses

First and foremost, we wish to evaluate how *mSARSA* performs in the environment detailed in *Subsection 4.2*. According to past research (Collenette et al.,

Table 4: Parameters used for the second section of the *Payoff Schema* experiment. Each column provides the payoff matrix utilitsed by each condition. *K* is the value of the Cooperation Index for that payoff matrix column.

Outcome	Payoff Value Used								
T	52	52	52	52	52	52	52	52	52
R	51	51	51	51	51	51	51	51	51
P	50	50	48	49	49	48	44	43	33
S	42	47	42	47	48	47	42	42	32
K	*0.1*	*0.2*	*0.3*	*0.4*	*0.5*	*0.6*	*0.7*	*0.8*	*0.9*

2018a), it is likely with the increased and deterministic interaction rate that any increase in the proportion of mutual cooperations may be dampened in comparison with past research. Given the model's overall success in mobile scenarios (Collenette et.al., 2017b), we expect that increased mood will help raise proportions of mutual cooperation in our *own* environment (over those of standard *SARSA* counterparts) even if this increase is smaller than in mobile agent research. If we do observe increased proportions of mutual cooperation (in comparison with standard *SARSA*), we expect that the model will uphold similarly when tested with the parameters from Collenette et.al., (2017b). As no prior direct data is available to make a prediction, we hypothesise that when *mSARSA* is played directly against *SARSA* agents, it may encourage SARSA to attain more mutually cooperative outcomes, based on the mixed-strategy results of Collenette et.al. (2018). This may only be true, however, after a critical proportion of *mSARSA* agents are present in the environment; the *Strategy Proportion* experiment serves to evaluate this hypothesis and search for such value. Lastly, it has been shown in previous research with human participants in the IPD that increasing *K* appears to promote increased rates of cooperation (see Wrightsman et.al., 1972; Colman et.al., 2018)). We therefore predict that it is possible *mSARSA* may follow this trend, particularly given that *mSARSA* utilises actual payoff values more extensively than its competitors. The varying of actual payoff values (**Conditions 1 through 3** versus **Conditions 5 through 7**) will serve to evaluate if the *K* value is purely responsible for any behavioural changes, or if payoffs must be chosen carefully in terms of exact values also.

5 RESULTS

5.1 Interaction Structure

As the primary reference point for the analysis of these data, we first summarise the result data of the comparable conditions from Collenette et.al., (2017b).

Table 5: Proportions of Outcome Types (Mutual Defections and Mutual Cooperations, respectively) converged to with 99% confidence intervals in the mobile environment **of past work**, summarised from Collenette et.al., (2017b). **S** column contains SARSA's performance, taken from the *Agentstate* condition (for comparison with our own SARSA tests).

Mutual Defection

mA State Info	0	0.4	0.8	S
Stateless	48.6% ± 1.5	19.7% ± 0.8	1.4% ± 0.1	
Agentstate	49.8% ± 1.1	24.3% ± 0.7	6.0% ± 0.5	49.7% ± 1.1
Moodstate	48.4% ± 1.2	24.4% ± 0.6	6.6% ± 0.5	

Mutual Cooperation

mA State Info	0	0.4	0.8	S
Stateless	8.4% ± 0.7	29.7% ± 1.0	78.9% ± 0.5	
Agentstate	21.1% ± 0.7	32.1% ± 0.8	63.2% ± 1.2	21.1% ± 0.7
Moodstate	21.3% ± 0.6	31.4% ± 0.8	62.3% ± 1.1	

Table 6: Mean Proportions of Outcome Types (Mutual Defections and Mutual Cooperations, respectively) averaged over the last 1000 rounds in the static environment **of the current experiments**. **S** column contains SARSA's performance. * *The [Agentstate, 0] condition did not converge on a final stable behaviour on any of the experimental trials.*

Mutual Defection

mA State Info	0 (SD)	0.4 (SD)	0.8 (SD)	S (SD)
Stateless	85.13% (0.24)	49.18% (11.19)	27.54% (6.98)	
Agentstate	*81.86%* *(23.08)**	49.61% (12.90)	26.54% (6.83)	94.99% (23.40)
Moodstate	85.43% (22.52)	49.58% (12.39)	26.80% (6.63)	

Mutual Cooperation

mA State Info	0 (SD)	0.4 (SD)	0.8 (SD)	S (SD)
Stateless	0.55% (0.24)	9.33% (2.87)	22.38% (5.52)	
Agentstate	*2.89%* *(4.39)**	9.52% (3.44)	23.85% (6.29)	0.07% (0.04)
Moodstate	1.54% (2.99)	9.36% (3.36)	24.09% (6.44)	

Average proportion of mean cooperation (over the last 1000 rounds) in the moody learning population followed a similar trend as in the resultant original data, if greatly reduced in actual value. The influence of the *Stateless* condition does not seem to be present in the static environment results as in the original mobile data, aside from possibly influencing *Mutual Cooperations* when *mA* is 0. An increase in information to the learner generally produced a greater proportion of mutual cooperation in interactions - as did an increase in the value of *mA*. Mutual defections similarly decreased as more information became available and *mA* increased, reaching a point where it was only slightly higher than mutual cooperation. Importantly, the proportions of *all* outcome types were approximately equal when *mA*=0.8; this is observable in *Table 6* with the similar levels of both mutual defection

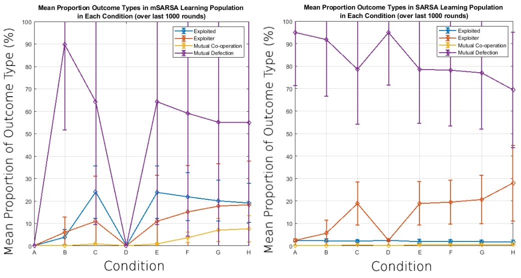

Figure 1: Summary graphs for *mSARSA* versus *SARSA*, displaying mean proportion of outcomes in the last 1000 rounds across the experimental conditions given in *Table 2*. Error bars show one standard deviation from the mean. Conditions A and D for *mSARSA* are marked as null as *mSARSA* agents were not present in the analysis population.

and mutual cooperation, which persisted through the non-mutual interactions also.

At all levels, *mSARSA* performed more mutual co-operations and less mutual defections than *SARSA*. It also showed a higher preference for mutual defection, and does not experience as many mutual cooperations, at all condition levels compared to the agents in Collenette et.al., (2017b). The results of a two-way between-subjects ANOVA on the *summed* mutual cooperation data from the last 1000 rounds of each experiment showed a significant effect of value of mA on the total amount of last 1000 rounds of mutual cooperation reached (F (2, 396) = 868.68, p < 0.01), and a significant (if weaker) effect of depth of state information on the same (F (2, 396) = 3.29, p = .0383). There no statistically significant interaction effect present between the two (at a significance level of p=.05).

5.2 Strategy Proportion

As the proportion of *mSARSA* increases in the environment, *SARSA*'s proportion of exploitation of its opponents increases. This increase is directly connected to its interactions with *mSARSA* agents, as evident by the low proportion of DC outcomes when the central *SARSA* agents are not in contact with any *mSARSA* agents (see **Conditions A** and **D**, where the analysed agents are not in direct contact with *mSARSA* partners). It maintains a high proportion of mutual defection throughout, as *SARSA* primarily defects in our environment. In the *mSARSA* population, the primary learned outcome types are mutual defection and Sucker outcomes. As the proportion of *mSARSA* in the network increases, mutual cooperation and Temptation outcomes increase - in line with the previous experiment where populations of entirely *mSARSA* have equal proportions of outcome types.

5.3 Payoff Schema

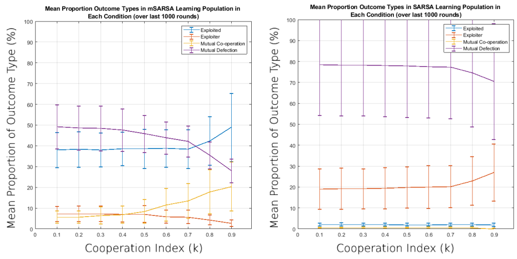

Figure 2: Summary graphs for *mSARSA* versus SARSA, displaying mean proportion of outcomes in the last 1000 rounds across the experimental conditions given in *Table 4*. Error bars show one standard deviation from the mean.

For the basic comparison of increasing values of *K* (whilst maintaining actual payoff values of the **T** and **R** payoffs), it seems that mutual cooperation rises in *mSARSA* agents (presumably with other *mSARSA* agents, as the *SARSA* outcome proportions show no such increase) as *K* gets larger. Mutual defections similarly decrease, and the proportion of **S** payoffs experienced also increases - meaning that overall, when *K* increases, *mSARSA* displays more cooperative behaviour to opponents. Outcomes **T** and **P** seem to be most influenced when *K*>=0.7, whereas those of **S** and **R** seem to change from approximately *K*=0.4.

With *SARSA*, it appears that behaviour is unaffected by the increase in *K*, except at high values (e.g.*K*>=0.7). This coincides with the point at which *mSARSA* begins to be exploited more often (presumably due to an increase in the latter's cooperation).

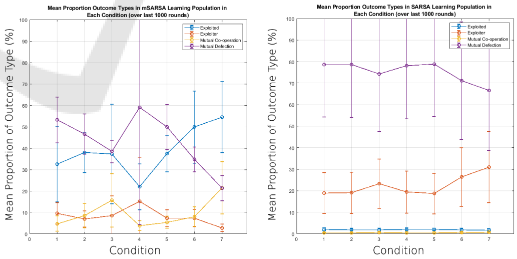

Figure 3: Summary graphs for *mSARSA* versus SARSA, displaying mean proportion of outcomes in the last 1000 rounds across the experimental conditions given in *Table 3*. Error bars show one standard deviation from the mean.

The change in outcome proportions experienced by the two strategies varies both between the smaller versus larger payoff values and as the value of *K* (the cooperation index) changes (*Figure 3*).

For *mSARSA*, it seems both of these manipulations to the payoff matrix has an influence on behaviour in certain ways. As the value of *K* increases, proportions of mutual defection decrease and proportions of Sucker outcomes increase. Mutual cooperations also increase, whereas proportions of Temptation outcomes do not appear to show a consistently linear pattern. Despite the classical payoff matrix having a *K*-value of 0.4, it does not seem to fit in-trend just under either of its *K*=0.5 counterparts (**Conditions 2** and **6**). This is possibly due to the size of the payoff values themselves in **Condition 4**, which are between 1 and 10 (as opposed to between 0 and 1, and 1 and 100 respectively for the other two categories).

When the actual value of payoffs are greater but the values of *K* follow the same trend (minimal, middling and high), these trends for *mSARSA* appear to be replicated but with greater exaggeration.

With *SARSA*, we see possible similar effects with the mutual defection and Temptation outcomes, but these differences may be purely reactive to the greater changes in *mSARSA* behaviour. As cooperative behaviour increases with payoff matrix change in *mSARSA*, *SARSA* observes more Temptation payoffs and slightly less mutual defections. As stated previously, this effect appears more exaggerated when the payoff values themselves are larger. *SARSA* seems fairly inoculated to changes in the payoff matrix in our environment, with only small amounts of variation between conditions and almost none at all in terms of mutually cooperative or Sucker outcomes. As *mSARSA* takes payoffs as a greater factor in each calculation (such as in the calculation of mood) however, it appears more influenced by its values and the ratio of cooperative payoffs to non-cooperative payoffs (*K*).

6 ANALYSIS, COMPARISON AND DISCUSSION

In Collenette et.al., (2017b), the primary trends of behaviour observed for *mSARSA* included an overall decrease in the proportion of mutual defections at convergence as the value of *mA* increased, with a likewise increase in mutual cooperations. The interaction of state information is a little more complex. The addition of more information above *Stateless* seems to keep mutual defections a little higher when *mA* gets larger, but still largely below the proportion that normal *SARSA* mutually defects at. Conversely, additional information aids mutual cooperation at lower levels of *mA*, but ultimately at the highest end (0.8) an increase in information hinders the proportion of mutual cooperations. Again, *mSARSA* still cooperates much more frequently than *SARSA*. When we transpose the algorithm into our environment, we observe a similar, if compressed, trend to that of the original data. The increased certainty and consistency of interactions in our environment may have led to more extreme proportions in both instances; for example, our *SARSA* defects at twice the proportion of the *SARSA* in Collenette et.al., (2017b), and mutually cooperates almost 400 times less (comparing averages). In terms of mutual defections, behaviour seemed largely unaltered by state information but followed the same trend as in Collenette et.al., (2017b) regarding *mA*.

Despite similar trends emerging to earlier work, however, the proportion of non-mutual outcomes gives us greater insight into how the algorithm functions here. Even though proportions of mutual defection decreased, and mutual cooperations increased, they evened out to similar proportions overall (25%) - as did non-mutual outcomes. This 'evening out' of all behaviours is unlike that observed in past work, where mutual cooperation in mSARSA was the clear dominant outcome. One possible explanation for these observations comes from how the value of mood changes within the algorithm in our environment. As covered in the final paragraph of *Subsection 3.1*, there are two possible equations to use for updating mood for each agent - MU1 (Collenette et al., 2017b) and MU2 (from later work; Collenette et.al., 2018). The latter of these has no direct data available for it (regarding mood changes or behaviour of the algorithm specifically), and so is impractical to compare against; it also exhibits cyclical behaviour of both mood and resulting behavioural outcomes, taking many more cycles to attempt to attain convergence (a primary limitation for our research in terms of time and hardware availability). It does, however, appear to provide *mSARSA* with the reactivity necessary for more diverse behaviour in such a high-interaction environment. We observed that with MU1, *mSARSA* agents rapidly increase in mood to maximum and then very rarely fluctuate from that position (if they do, the mood change is infinitesimally small). These effect may not have been observed in Collenette et.al., (2017b) due to the lower rate of interaction and the lack of guarantee of interacting with the same partner repeatedly. We opted not to utilise MU2 due to reasons of practicality (after initial testing), but this is certainly an avenue of future research worth pursuing.

In essence, agents in our own experiments were therefore always operating at extremely high moods, meaning that when any agent initially selects defection, a new behaviour is then re-selected under the new, higher ε value (0.9), which is high enough to lead to near-random behaviour selection. Data not reported in *Table 6* is that of the non-mutual behaviours observed, which were at similar levels to both the mutual outcomes - suggesting that behaviour is more randomly selected during the ‘converged’ portion of the experiment (at the very least, within the last 1000 rounds of each test). This is of important note to future researchers who opt to use this algorithm, as an uninformed selection of interaction structure and environment may lead to differing efficacy.

Effects of increasing the proportion of *mSARSA* in the environment are relatively straight-forward. As *mSARSA* has a comparatively high propensity to cooperate, increasing the proportion of *mSARSA* agents in the environment increases the likelihood it will be exposed to agents of its own kind - this explains the gradual increase of the mutually cooperative and exploited outcome proportions (the **R** and **S** payoffs respectively). We can see this is the case from looking at the *SARSA* data; in conditions where the analysed central group of agents were all made up of *SARSA* agents that did not come into direct contact with *mSARSA* agents (i.e. conditions *A* and *D*, where *mSARSA* agents were in the outer corners of the grid). In these conditions, *SARSA*’s proportion of **T** payoffs is as minimal as its other payoffs (as it primarily plays with like agents, who also defect heavily). As the proportion of more cooperative (*mSARSA*) opponents increases, it finds more opportunities to exploit. When even 10% of grid agents utilise SARSA, this significantly impacts the outcomes experienced by *mSARSA*; they experience twice as many mutual defections when there are some *SARSA* agents present in the environment than when there were none (compared with data from *Interaction Structure*). This may suggest that *mSARSA* is particularly sensitive, in this environment, to highly defective or pure-defective partners. Our data also suggest that in our experiments there was no migration of behaviour over time (at least within the time frame of 10,000 rounds) - in contrast to our hypotheses, the introduction of *mSARSA* to an environment of *SARSA* did not encourage the latter to cooperate more.

One of the more interesting outcomes of the three experiments dictated here is the interaction of *mSARSA* with *K*, the cooperation index value. Where we observed approximately 60% mutual defections and 5% mutual cooperations in *Strategy Proportion:* **Condition F**, when 48% of agents in the environment were *mSARSA*, the alteration of the payoff scheme towards a higher *K* facilitates half as many mutual defections and four times as many mutual cooperations. The standard payoff matrix (used throughout *Strategy Proportion* tests) has a *K* value of 0.4, meaning that the value of the payoffs themselves only appears to have decreased the **T** proportion slightly, and increased the R proportion slightly (in *Figure 2* from in *Figure 1*). Observing the consistent trend in human research that an increased *K* leads to increased rates of cooperation (Wrightsman et.al., 1972; Colman et.al., 2018), it is very positive to see that the more human-like of the two learning algorithms reacts similarly. The choice of payoff matrix therefore is an important tool in maximising mutual cooperation in the network outlined in this research. We also observe an effect of the size of the actual payoffs themselves on the outcome behaviours observed, which is critical information for future experimentation. *mSARSA* seems to display larger changes to its mutual outcomes when *K* increases and the actual value of the payoffs themselves are large (see *Conditions 5, 6, 7* in *Figure 3*).

7 CONCLUSIONS

Overall, the experiments detailed have thoroughly tested the *mSARSA* algorithm presented in Collenette et.al. (2017b) and Collenette et.al. (2018) with a novel interaction structure and environment, and also tested two further dimensions of interest (environmental presence and cooperation index). We observed similar trends to these past works, but also highlight the limitations of differing versions of the algorithm within our grid network where possible. Namely, there was an exaggeration of weaknesses in the first version of the algorithm, which were altered by the authors in later work but in ways that possibly present further technical issue. The moody alternative to *SARSA* cooperates more than its standard counterpart, but is vulnerable to exploitation in its first iteration, meaning that in this particularly intense interaction structure it does not adapt well to poor outcomes. Additionally, whilst increasing the proportion of *SARSA* agents in the environment serves to reaffirm this vulnerability to exploitation, altering the cooperation index *K* demonstrated that *mSARSA* follows behavioural reactivity to this variable previously observed in humans. This offers affirmation to the design basis of the variant algorithm, and creates potential for more accurate human behavioural modelling in future.

We would next like to perform similar evaluations with the second iteration of the algorithm and repeat experiments with *K* with differing levels of agent connectivity to test for interaction effects. We are currently evaluating the algorithm in dynamic random networks as an extension of this. There are also many avenues of experimentation possible for altered versions of *mSARSA*; using alternative methods to evaluate the desirability of an outcome, for example.

ACKNOWLEDGEMENTS

This work was part-funded by UK Research and Innovation (UKRI) and Loughborough University. We would like to thank Dr. Joe Collenette and Dr. Katie Atkinson (University of Liverpool) for their provision of the code from their implementation of Moody SARSA, and for their friendly support and communication.

REFERENCES

Barreda-Tarrazona, I., Jaramillo-Gutiérrez, A., Pavan, M., and Sabater-Grande, G. (2017). Individual characteristics vs. experience: An experimental study on cooperation in prisoner's dilemma. *Frontiers in Psychology*, 8.

Belkaid, M., Cuperlier, N., and Gaussier, P. (2017). Emotional metacontrol of attention: Top-down modulation of sensorimotor processes in a robotic visual search task. *PLoS ONE*, 12(9).

Clore, G. L. and Ortony, A. (2013). Psychological construction in the occ model of emotion. *Emotion Review*, 5(4):335 – 343.

Collenette, J., Atkinson, K., Bloembergen, D., and Tuyls, K. (2017a). Environmental effects on simulated emotional and moody agents. *The Knowledge Engineering Review*, 32.

Collenette, J., Atkinson, K., Bloembergen, D., and Tuyls, K. (2017b). Mood modelling within reinforcement learning. In *Proceedings of ECAL'17*, pages 106–113. MIT Press.

Collenette, J., Atkinson, K., Bloembergen, D., and Tuyls, K. (2018a). Modelling mood in co-operative emotional agents. In *Distributed Autonomous Robotic Systems*, volume 6.

Collenette, J., Atkinson, K., Bloembergen, D., and Tuyls, K. (2018b). On the role of mobility and interaction topologies in social dilemmas. In *Proceedings of Conference on Artificial Life*.

Colman, A. M., Pulford, B. D., and Krockow, E. M. (2018). Persistent cooperation and gender differences in repeated prisoner's dilemma games: Some things never change. *Acta psychologica*, 187:1–8.

Fehr, E. and Schmidt, K. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, pages 817–868.

Fudenberg, D., Rand, D. G., and Dreber, A. (2012). Slow to anger and fast to forgive: Cooperation in an uncertain world. *Law & Prosociality eJournal*.

Gao, Y. (2012). A reinforcement learning based strategy for the double-game prisoner's dilemma. In *Proceedings of the First International Conference on Agreement Technologies*, volume 918, pages 317–331.

Hertel, G., Neuhof, J., Theuer, T., and Kerr, N. L. (2000). Mood effects on cooperation in small groups: Does positive mood simply lead to more cooperation? *Cognition and Emotion*, 14:441 – 472.

Hristova, E. and Grinberg, M. (2005). Investigation of context effects in iterated prisoner's dilemma game. In *CONTEX. Lecture Notes in Computer Science*, volume 3554, Berlin, Heidelberg. Springer.

Lin, B., Cecchi, G. A., Bouneffouf, D., Reinen, J., and Rish, I. (2019). Reinforcement learning models of human behavior: Reward processing in mental disorders. *ArXiv*, abs/1906.11286.

Ortony, A., Clore, G. L., and Collins, A. M. (1988). The cognitive structure of emotions. New York, NY: Cambridge University Press.

Palfai, T. P. and Salovey, P. (1993). The influence of depressed and elated mood on deductive and inductive reasoning. *Imagination, Cognition and Personality*, 13:57 – 71.

Proto, E., Sgroi, D., and Nazneen, M. (2017). The effect of positive mood on cooperation in repeated interaction. In *The Warwick Economics Research Paper Series (TWERPS) 1141*, University of Warwick, Department of Economics.

Shteingart, H. and Loewenstein, Y. (2014). Reinforcement learning and human behavior. *Current Opinion in Neurobiology*, 25:93–98.

Sutton, R. and Barto, A. (2018). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, Mass.

Team, P. M. (2021). projectmesa/mesa. https://github.com/projectmesa/mesa. Last accessed 03 August 2021.

Vaughan, R. T. (2008). Massively multi-robot simulation in stage. *Swarm Intelligence*, 2:189–208.

Wilson, A. J. and Wu, H. (2017). At-will relationships: How an option to walk away affects cooperation and efficiency. *Games Econ. Behav.*, 102:487–507.

Wrightsman, L. S., O'Connor, J., and Baker, N. J. (1972). *Cooperation and Competition: Readings on Mixed-Motive Games*. Brooks/Cole Pub. Co., Belmont, CA.

Yu, C., Zhang, M., Ren, F., and Tan, G. (2015). Emotional multiagent reinforcement learning in spatial social dilemmas. *IEEE Transactions on Neural Networks and Learning Systems*, 26:3083–3096.

Zulfiqar, S. and Islam, A. (2017). Exploring the role of emotions and moods in decision making: Study on the use of structured decision approach and intuition. *International Journal of Engineering and Management Sciences*, 2:140–149.