Mood Modelling within Reinforcement Learning

Joe Collenette[1], Katie Atkinson[1], Daan Bloembergen[1] and Karl Tuyls[1]

[1] Department of Computer Science, University of Liverpool, UK
j.m.collenette@liverpool.ac.uk

Abstract

Simulating mood within a decision making process has been shown to allow cooperation to occur within the Prisoner's Dilemma. In this paper we propose how to integrate a mood model into the classical reinforcement learning algorithm Sarsa, and show how this addition can allow self-interested agents to be successful within a multi agent environment. The human-inspired moody agent will learn to cooperate in social dilemmas without the use of punishments or other external incentives. We use both the Prisoner's Dilemma and the Stag Hunt as our dilemmas. We show that the model provides improvements in both individual payoffs and levels of cooperation within the system when compared to the standard Sarsa model. We also show that the agents' interaction model and their ability to differentiate between opponents influences how the reinforcement learning process converges.

Introduction

This paper explores how groups of self-interested agents can be designed in a way that allows cooperation to develop over time in social dilemmas. The agents should be able to develop cooperation without the use of external interference, such as punishments for defection and incentives for cooperation. We have chosen to design these agents in a human-inspired way using mood to influence their learning. We make use of reinforcement learning techniques to show that using a human-inspired computational model of mood allows self-interested agents to develop cooperation with other agents over time.

In particular we investigate whether the mood model developed by Collenette et al. (2016b) can be incorporated within the classical reinforcement learning algorithm Sarsa, and whether this will allow agents to converge towards cooperation in a social dilemma. Additionally we consider different types of interaction between the agents, depending on whether or not they can distinguish their opponent and whether they can observe their opponent's current mood. We will be exploring this model in both the Prisoner's Dilemma and the Stag Hunt social dilemmas, noting the differences our model gives to these different dilemmas.

It is worth noting that emotions and mood are two distinct aspects of the human psyche. Emotions are short-term feelings that are directed to an individual object or person, which can change quickly (Levenson, 1994). Mood is the reverse of this, as it does not change quickly and is undirected. It is a general feeling that an individual has (Gray et al., 2001). It has been shown that both emotion and mood have an effect on decision making in humans (Schwarz, 2000; Hertel et al., 2000). While we acknowledge that emotions and mood have physiological effects as well (Keltner and Gross, 1999), within the scope of this paper we incorporate only the psychological effects mood has on learning and decision making.

While there has been research into how emotions can be integrated into agents (Collenette et al., 2016a; Lloyd-Kelly et al., 2014) and how they can be represented there has been little work on how mood can be used. Mood has often been used as a black box (Ojha and Williams, 2016; Santos et al., 2009) where the mood does not affect the decision making directly, but affects other aspects of the agent. In contrast here we set out a clear definition of how mood is incorporated into decision making, with grounding in psychology. We focus on the mood as the main aspect of this work.

Mood Model

The mood model of Collenette et al. (2016b) provides a generic framework of mood, grounded in psychology, which can be incorporated into other processes. Mood is represented as a real number in the range of 0 to 100 where low values represent negative moods and high values represent positive moods. This spectrum ranging from low to high reflects how psychologists view mood as well (Hepburn and Eysenck, 1989). The numbers of 0 and 100 were chosen to give an intuitive understanding of where the mood lies and for easier integration with other parts of the mood model framework. The framework uses the Homo Egualis model of fairness (Fehr and Schmidt, 1999) as a basis to define how an agent's perception of their payoffs relative to other agents affects their mood.

Within a social scenario after an interaction has occurred between two agents we will update each of their mood values. How the mood is updated is shown in Definition 1.

Carole Knibbe et al, eds., Proceedings of the ECAL 2017, Lyon, France, 4-8 September 2017,
(Cambridge, MA: The MIT Press, ©2017 Massachusetts Institute of Technology).

Definition 1 (Mood Calculation) *Let AG be the set of all agents, with i and $j \in AG$. Let t denote time. Let p_i^t return the payoff of agent i at time t. Let m_i^t return the mood of agent i at time t, in the range $0 < m < 100$. Let μ_i^t denote the average payoff for agent i up to time t. Let F_i^t return the opponent of agent i at time t. Let $\alpha = \beta$ be in the range 0 to 1.*

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

$$\Omega_{i,j}^t = \mu_i^t - \alpha_i^t \cdot \max(\mu_j^t - \mu_i^t, 0) - \beta_i^t \cdot \max(\mu_i^t - \mu_j^t, 0) \quad (2)$$

$$m_i^t = m_i^{t-1} + (p_i^t - \mu_i^{t-1}) + \Omega_{i,j}^{t-1} \text{ where } j = F_i^t \quad (3)$$

Equation 1 gives us the mood of the agent that can be used in the Homo Egualis equation. It gives high moods as a low number and low moods as a high number. For example a mood of 75 will give a α of 0.25. Equation 2 is a two agent version of the Homo Egualis equation, which returns an adjusted payoff based on how the agent views its opponent based on a comparison of rewards. We are using average payoffs rather than total payoffs, as in our scenarios there is no guarantee that the number of games played will be the same. In the Homo Egualis equation we take $\alpha = \beta$, representing an idealistic view in which an agent views other agents equally to itself.

Finally Equation 3 gives the adjustment to the mood value by taking the difference between the average payoff and the received payoff with the Homo Egualis adjustment.

Social Dilemmas

Our experimental settings will be the Prisoner's Dilemma and the Stag Hunt. In these social dilemmas, two players have a choice of cooperation or defection where the decision is made simultaneously, without prior communication on how the agents will behave. The payoff matrices we will use for the Prisoner's Dilemma and the Stag Hunt are given in Table 1, where cooperation is given as C, defect is given as D, and the numbers indicate the payoff each player gets when choosing the indicated actions.

C, C	D, D	C, D	D, C
$3, 3$ $(3, 3)$	$1, 1$ $(1, 1)$	$0, 5$ $(0, 2)$	$5, 0$ $(2, 0)$

Table 1: Payoff matrix of the Prisoner's Dilemma and the Stag Hunt. The Stag Hunt is shown in parentheses

It is in the best interests of both agents to cooperate in both dilemmas as this will give the best result for the group as a whole. However there is an incentive to defect in the Prisoner's Dilemma as this leads to a higher payoff for the defector at the expense of the cooperator. If both agents reason this way in the Prisoner's Dilemma then the group as a whole gets the worst payoff as they both defect, highlighting the dilemma of the game and giving rise to the Nash Equilibrium of mutual defection.

When compared to the Prisoner's Dilemma the Stag Hunt shares many dynamics, but the main change is that there are now two Nash Equilibriums; these are mutual cooperation and mutual defection. The mutual defection Nash is a risk-dominant strategy as when the opponent deviates, there is no risk of reducing the payoff. However agents can increase their own payoff and the group's payoff by choosing to cooperate; this is a risky move as if the other agent does not reciprocate the cooperation, the payoff is lost.

Exploring how to encourage cooperation to evolve within these groups of self-interested agents, has been an active topic of research (Axelrod and Hamilton, 1981; Santos et al., 2008; Bloembergen et al., 2014; Skyrms, 2004; Bolton et al., 2016). It is for this reason that we adopt this model of interaction in the current work as well.

Reinforcement Learning

Reinforcement learning (Sutton and Barto, 1998) prescribes how an agent can learn to optimise her behaviour by repeated trial-and-error interaction with the environment. At each time step, the agent takes an action based on the current state of the environment, and observes its effect in terms of a reward signal and resulting state change. Behaviour that yield high rewards will be reinforced, whereas behaviour that causes low rewards or penalties will be reduced. The goal of the learning agent is to maximise her expected future rewards.

One of the most well-known reinforcement learning algorithms is Sarsa. Sarsa learns state-action values, $Q(s_t, a_t)$, which represent the expected sum of (discounted) future rewards after taking action a in state s at time step t. Definition 2 gives the Q update function for Sarsa given the immediate reward r_{t+1} and the expected future rewards, estimated recursively by the value of the next state-action pair $Q(s_{t+1}, a_{t+1})$ and discounted by γ.

Definition 2 (Sarsa (Sutton and Barto, 1998).) *Let S be the set of states with $s \in S$. Let A be the set of actions with $a \in A$. Let t be the time, r represent the reward, α the learning step size and γ the discount factor of future rewards. Then, Sarsa updates $Q(s_t, a_t)$ using the following equation:*

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$$

Reinforcement learning has been applied to social dilemmas, exploring different aspects of the problem, such as different variations of techniques (Masuda and Ohtsuki, 2009), and how they can be tweaked to allow greater cooperation against a variety of opponents (Vassiliades et al., 2011; Crandall and Goodrich, 2005). There has been exploration of how different interaction networks can affect the spread of cooperation (Ranjbar-Sahraei et al., 2014), as well applying emotions to the reinforcement model (Yu et al., 2015).

Carole Knibbe et al, eds., Proceedings of the ECAL 2017, Lyon, France, 4-8 September 2017,
(Cambridge, MA: The MIT Press, ©2017 Massachusetts Institute of Technology).

The majority of the work completed in this area has focused on agents which are static or are represented in a dynamic network. In our implementation we simulate mobile agents randomly navigating an environment, with no guarantees on the number of interactions and who interacts with who.

In our experiments we build on and compare to the standard implementation of the Sarsa algorithm, which is known as an on-policy algorithm as it uses it's current policy as a predictor for future rewards, through $Q(s_{t+1}, a_{t+1})$ (Q-learning in contrast uses a greedy policy as predictor, $\max_{a'} Q(s_{t+1}, a')$). We choose an on-policy method because mood by definition is an on-policy adaptation since we learn through experiences and adapt as the environment changes through the use of mood (Rinck et al., 1992).

Mood Model Integration

To integrate the mood model (Definition 1) into the Sarsa algorithm (Definition 2) we change the way in which the estimation of future rewards is computed. Where Sarsa uses $Q(s_{t+1}, a_{t+1})$, the value of the next state-action pair, as an estimate, we replace this by an estimate Ψ based on mood, yielding the update rule

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

In our experiments we set $\alpha = 0.1$ and $\gamma = 0.95$.

Our estimation Ψ uses a memory which stores the payoffs an agent receives in that state action pair. For example if the agents are able to distinguish between opponents then after a number of interactions the agent will have stored the results it received when playing that opponent with each action. To calculate the estimation we use the mean of the previous interactions for that action and state. The mood is integrated by changing how far back the agent will look to provide the mean. For example if the mood is 25 then the agent will look back at 75% of the previous outcomes and calculate the mean payoff received. The computation is formalised in Definition 3. It reflects how people in bad moods will think longer about a problem compared to those in a good mood, who use a more instinctive response (Hertel et al., 2000). We limit the memory to 20 interactions as a longer history will reduce the effect that an individual interaction has on the average. However, the effect of an individual interaction should also not be too extreme. By setting the maximum to 20, we find a balance where the effect is preserved while a single interaction will not override past experience.

Definition 3 (Estimation of Future Rewards) *Let Mem_i^a be the set of rewards obtained by agent i when using action a where $|Mem_i^a|$ is at maximum 20, and $Mem_i^a(0)$ returns the most recent reward. Let m_i return the mood of agent i.*

$$\alpha_i = (100 - m_i)/100 \tag{4}$$

$$\beta = ceil(|Mem_i^a|/\alpha_i) \tag{5}$$

$$\Psi = \left(n \sum_{0}^{\beta} Mem_i^a(n) \right) / \beta \tag{6}$$

To choose which action to take, we use the ϵ-greedy method. This method selects the action with the highest Q value with probability $1 - \epsilon$, and a random other action with probability ϵ. In our implementation we set $\epsilon = 0.1$ initially. We have also integrated the mood model into how ϵ is chosen. In neutral moods there is no change in the value of epsilon. If the agent is in a bad mood ($m < 30$) and has chosen to cooperate we increase ϵ to reflect how humans are more likely to defect in these kinds of social dilemmas (Haley and Strickland, 1986). When the agent is in a good mood ($m > 70$) then we will likewise increase ϵ if they choose to defect. We are reflecting how people in good moods are more likely to choose an idealist option, even if that choice is risky, as discussed in (Hertel et al., 2000; Leahy, 2005).

Finally, an important choice when applying reinforcement learning is how to initialise the state-action values $Q(s, a)$. In our experiments we use the first reward that the agent receives for that state action pair as the initial value, as this best reflects how people learn about new experiences. For example, Shteingart et al. (2013) show that when resetting the initial values to new data Q-learning can predict how people choose between a risky option and a safe option.

We note that the mood model shares some similarities with reward shaping. Reward shaping supplies an additional reward to the reward that would normally be received for a particular action (Ng et al., 1999). For the Sarsa algorithm the update rule will be updated to the following.

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + F(s_t, a_t, s_{t+1}) + \\ + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$$

Where F returns a seperate reward based on the state action pair and the next state. The reward is defined by the designer of the system. This technique can allow reinforcement learning to scale up to more complex problems (Devlin et al., 2011; Randløv and Alstrøm, 1998). The main difference between reward shaping and the mood model is that the shaping reward function supplies additional rewards, whereas our mood model directly affects decision making and the estimation of future rewards.

Experimental Set-up

The experiment will be conducted using the stage library (Vaughan, 2008) simulating 70 agents within an environment, shown in Figure 1.

The agents will engage in a random walk around the arena. When an agent has line of sight and is sufficiently close to another agent then those two agents will engage in an iteration of either the Prisoner's Dilemma or the Stag Hunt game depending on the scenario. Then they will both continue moving around the environment. This will continue for 10 minutes, then the arena will reset (to prevent

Scenario	1	2	3	4	5	6	7
MA	0	0.2	0.4	0.6	0.8	v	Sarsa

Table 2: Scenarios with different values of MA.

agents getting stuck in corners of the environment) with the agents retaining their memory and Q values. This is repeated until they converge. Convergence is said to have occurred when the average proportions of outcomes from the last 5 10 minute runs are within 0.005 of the average proportions of the last 25 10 minute runs.

Scenarios

We simulate across different scenarios, which include variations on how much the mood value affects the choices made by the agents (which we indicate as MA) and can be seen in Table 2. We start with no effect so that we can compare how the addition of memory alone affects the choices made. We will then start to increase the value of MA to see the point at which the reinforcement learning converges to cooperation. In scenario 6, the value of MA is a variable amount v which is dependant on the mood, the exact calculation of v is shown in Definition 4. Finally we will compare this against standard *Sarsa* to analyse the effect of our implementation.

For each of these scenarios compare three different definitions of the state space. The first is *Stateless*, where the agents have no knowledge of their environment or who they are interacting with. The second is *Agent State*, where the agents can distinguish between the opponents they are interacting with. Finally, the *Mood State* is where the agents additionally observe the mood their opponent is in. The definitions are given in Table 3 where AG is the set of agents and $MV = \{High, Neutral, Low\}$ is the set of mood representations. Given the mood m_i of agent i, MV_i is *High* when $m_i > 70$, *Low* when $m_i < 30$, and *Neutral* otherwise.

Definition 4 (Variable MA Value) *Let m_i^t return the mood value of agent i at time t.*

$$v = \begin{cases} (m_i^t - 50)/100 & \text{if } m_i^t > 70 \\ (50 - m_i^t)/100 & \text{if } m_i^t < 30 \\ 0.1 & \text{otherwise.} \end{cases} \quad (7)$$

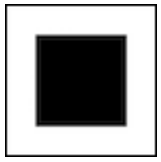

Figure 1: The simulated environment. White areas are traversable. Measures $5m^2$ with an interior block of $4m^2$,

Name	S	Description
Stateless	$\emptyset$	No state information used
Agent State	AG	Agents can distinguish between opponents
Mood State	$AG \times MV$	Agents can distinguish between opponents and observe their mood

Table 3: State definitions

Hypotheses

We predict that the memory only experiment will show a small amount of variation when compared to the Sarsa algorithm as the only difference between them is how they predict future rewards (H1). This leads to the same actions being selected in both cases. In both of these cases we predict that in both types of social dilemma they will converge to defection in a stateless scenario as the opponents they will be facing will be randomised, preventing any cooperation from being sustained. When the agents are able to distinguish between opponents we predict that some small levels of cooperation will appear that will then fade into defection as the decision making agent may randomly switch due to the ϵ greedy action choice (H2).

In regards to the other scenarios we predict that high levels of MA will give higher levels of cooperation (H3); this is due to an effect with the mood model that Collenette et al. (2016b) showed. The levels of mood will increase with mutual defection, however when the agent starts to cooperate the mood will fall, causing the other agents' mood to go up. The rate at which the mood decreases in the original agent will be less than the rate at which it rises in the opponent, ending with the overall mood increasing, which in turn will lead to the agents using the higher ϵ when choosing to defect. H1, H2, and H3 refer to our hypotheses.

Results and Analysis

Understanding whether the introduction of this mood model has been a success we first need to define what success means. We look at two criteria for success, firstly we look at whether mutual cooperation can be created and sustained, secondly we look at the average payoffs for the agents to see if the agents are better off using our mood model as compared to Sarsa.

Tables 4, 5, and 6 show the proportions of the different outcomes the agents converged to, along with their 99% confidence values for the *Stateless*, *Agent State*, and *Mood State* scenarios respectively.

We first note that while Sarsa has shown a strong preference for cooperation in prior work (Vassiliades et al., 2011), in our work it shows a strong preference for defection. This is due to a change from a two agent setting to a larger setting, which is reflected in the cooperation increasing when

Carole Knibbe et al, eds., Proceedings of the ECAL 2017, Lyon, France, 4-8 September 2017,
(Cambridge, MA: The MIT Press, ©2017 Massachusetts Institute of Technology).

S	G	Coop	Defect	Non Mutual
1	PD	0.084±0.007	0.486±0.015	0.431±0.011
1	SH	0.114±0.007	0.436±0.015	0.450±0.011
2	PD	0.144±0.009	0.386±0.014	0.470±0.010
2	SH	0.157±0.009	0.369±0.013	0.474±0.009
3	PD	0.297±0.010	0.197±0.008	0.506±0.008
3	SH	0.313±0.011	0.207±0.010	0.481±0.008
4	PD	0.499±0.010	0.089±0.004	0.412±0.008
4	SH	0.804±0.006	0.010±0.001	0.186±0.006
5	PD	0.789±0.005	0.014±0.001	0.197±0.005
5	SH	0.809±0.004	0.010±0.001	0.181±0.004
6	PD	0.366±0.008	0.149±0.006	0.484±0.006
6	SH	0.634±0.026	0.060±0.009	0.306±0.019
7	PD	0.017±0.002	0.765±0.007	0.218±0.007
7	SH	0.028±0.006	0.735±0.018	0.236±0.012

Table 4: Proportions of outcomes converged to with 99% confidence intervals, for each Scenario (S) and each Game (G), when no state information is used (*Stateless*).

S	G	Coop	Defect	Non Mutual
1	PD	0.211±0.007	0.498±0.011	0.290±0.011
1	SH	0.216±0.007	0.499±0.010	0.285±0.009
2	PD	0.247±0.009	0.393±0.010	0.361±0.008
2	SH	0.230±0.007	0.384±0.011	0.387±0.010
3	PD	0.321±0.008	0.243±0.007	0.436±0.007
3	SH	0.338±0.008	0.222±0.007	0.440±0.008
4	PD	0.427±0.009	0.143±0.006	0.431±0.007
4	SH	0.463±0.010	0.123±0.006	0.414±0.009
5	PD	0.632±0.012	0.060±0.005	0.308±0.009
5	SH	0.632±0.011	0.058±0.004	0.310±0.009
6	PD	0.361±0.009	0.193±0.007	0.446±0.007
6	SH	0.376±0.010	0.190±0.007	0.433±0.007
7	PD	0.211±0.007	0.497±0.011	0.293±0.009
7	SH	0.220±0.007	0.494±0.011	0.286±0.010

Table 5: Proportions of outcomes converged to with 99% confidence intervals, for each Scenario (S) and each Game (G), using the *Agent State*.

the agents are able to distinguish between different opponents. While this explains some of the differences, we also note that there are effects from allowing our agents to move and the resulting inconsistency in the number of interactions. Introducing movement into our scenarios also introduces randomness into when any two agents may interact. This randomness does not allow accurate predictions on who the next opponent will be, or whether any two particular agents will converge in their pairwise interactions.

By comparing the Sarsa (scenario 7) outcomes to our memory only outcomes (scenario 1), we note that there is a small improvement to the memory only outcomes when the agents are anonymous, which is in contrast with our hypothesis H1. The difference in the improvements is down to how Sarsa predicts future outcomes based on its current Q value, which takes into account all previous interactions, whereas the mood agents use a limited memory of recent outcomes which allows it to adapt to the prevailing action consensus quicker than the Sarsa agents.

Similarly, for the scenarios we tested, the addition of states allows cooperation to increase, while introducing larger amounts of non mutual actions and reducing mutual defection. This is in contrast to our hypothesis H2 which stated that any cooperation created would fall. However there was an exception with larger values of MA, where cooperation decreases with the addition of states. The addition of states increases the instability of the system, if the value of MA is high then the chance of an agent defecting is reduced to the point where cooperation spreads more effectively than defection. The reduction of information allows this cooperation to spread as agents try to converge on the group of agents as a whole rather than on an individual level. When MA values are low, then the additional informations helps to prevent the spread of defection.

Hypothesis H3 was confirmed as higher levels of MA increase the level of cooperation. The differences between the level of cooperation in the Stag Hunt and Prisoner's Dilemma show that lower levels of MA are required, to yield a high proportion of cooperation in the Stag Hunt. The difference is due to the payoff structure (Table 1), as in the Stag Hunt mutual cooperation gives a higher payoff than the non mutual action does for the defector. The value of MA gives a higher guarantee than cooperation will be mutual so it is in the interest of the agent to cooperate, which is reflected in their Q Values, whereas in the Prisoner's Dilemma the temptation to defect is still there as the individual payoff for defecting is higher than mutual cooperation.

Finally we can see that we can achieve high levels of cooperation in the Stag Hunt and a majority of cooperation in the Prisoner's Dilemma. Next we will be looking at the second way to measure success, which is through the payoffs received by the agents. Figures 2, 3, and 4 show the average score of an agent through each run, for each type of state definition. We have chosen average scores rather than total scores as the number of interactions per run is not consistent across agents. The number of runs is different as the time to convergence is different. We look at differences between scenarios 7 (Sarsa, Figure 2), 1 (Memory only, Figure 3), and 6 (Variable Mood, Figure 4) in the Prisoner's Dilemma.

We can see from the three figures that the most successful agents were in the mood scenario, which is reflected by the higher levels of cooperation as noted previously. There are only minor differences between the different types of state. An exception is shown in the stateless scenario for Sarsa, where the average score is much lower when compared to the agent and mood states, which is reflected by the lower

Carole Knibbe et al, eds., Proceedings of the ECAL 2017, Lyon, France, 4-8 September 2017,
(Cambridge, MA: The MIT Press, ©2017 Massachusetts Institute of Technology).

S	G	Coop	Defect	Non Mutual
1	PD	0.213±0.006	0.484±0.012	0.303±0.010
1	SH	0.216±0.008	0.481±0.011	0.302±0.009
2	PD	0.246±0.007	0.379±0.008	0.375±0.008
2	SH	0.236±0.007	0.384±0.010	0.380±0.008
3	PD	0.314±0.008	0.244±0.006	0.442±0.008
3	SH	0.319±0.008	0.234±0.005	0.447±0.007
4	PD	0.454±0.009	0.135±0.007	0.411±0.007
4	SH	0.481±0.011	0.110±0.006	0.409±0.009
5	PD	0.623±0.011	0.066±0.005	0.311±0.007
5	SH	0.627±0.012	0.061±0.005	0.312±0.009
6	PD	0.365±0.008	0.194±0.005	0.441±0.008
6	SH	0.371±0.009	0.189±0.006	0.440±0.008
7	PD	0.211±0.007	0.483±0.012	0.306±0.010
7	SH	0.213±0.007	0.495±0.011	0.292±0.010

Table 6: Proportions of outcomes converged to with 99% confidence intervals, for each Scenario (S) and each Game (G), using the *Mood State*.

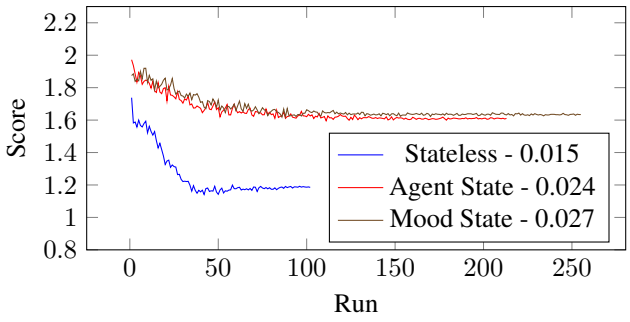

Figure 2: Average scores for Sarsa in the Prisoner's Dilemma over each run, with 99% confidence interval of final average score.

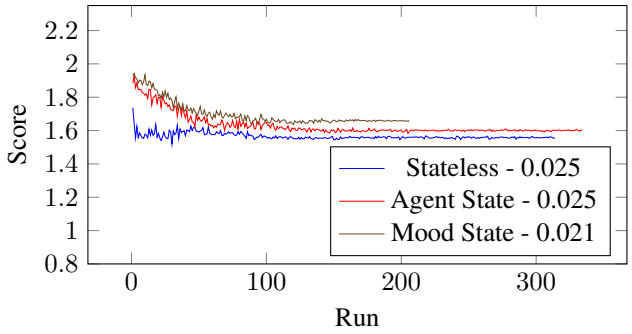

Figure 3: Average scores for memory only in the Prisoner's Dilemma over each run, with 99% confidence interval of final average score.

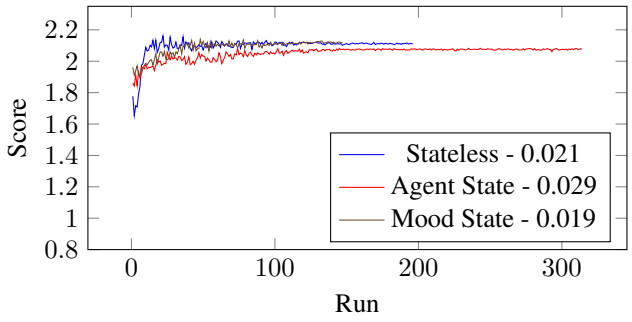

Figure 4: Average scores for variable mood in the Prisoner's Dilemma over each run, with 99% confidence interval of final average score.

cooperation levels.

When the game is changed to the Stag Hunt as shown in Figures 5, 6, 7, we see that the same scenarios are successful. Figure 7 shows an exception, here stateless is the most successful by a large margin. To see why this is the case we need to look at the form of the games themselves, how actions are chosen, and how they converge for our agents.

The main difference in individual payoffs between the Stag Hunt and the Prisoner's Dilemma is that in the Stag Hunt the payoff for defecting against a cooperating agent is lower than the individual payoff for mutual cooperation. In the Prisoner's Dilemma this individual payoff is higher when defecting against a cooperating opponent. A perfectly rational agent will therefore choose to cooperate in the Stag Hunt if they know the other opponent is cooperating. However in our scenario choosing cooperation is not guaranteed by the ϵ-greedy choice; when an agent inadvertently defects they receive the temptation payoff. In the Prisoner's Dilemma the agent will then continue defecting which leads to the drop in average score as more agents choose defection, as their Q Values reflect the higher value that defection brings. In the Stag Hunt, agents will continue to choose cooperation as the Q Value for defection does not rise higher than the Q Value for cooperation.

When the agents can differentiate between opponents, the agent that receives the defect payoff will respond to the defecting agent at an individual level, rather than defecting against all other opponents. This prevents the spread of defection in the Prisoner's Dilemma. This is shown by the average scores being higher in the agent and mood states when compared to the stateless in the memory only and Sarsa scenarios. However in the Stag Hunt the cooperation has not spread for the same reason: cooperation needs to be created on a individual level first, which is shown by the weaker agent state and mood state average payoffs.

When using the mood scenario we see that stateless performs just as well as the agent and mood states in contrast to the memory only and Sarsa scenarios. The reason that defection does not spread is due to the effects mood has on action selection. There is a initial drop in average payoff as defection spreads, however as the mood is going up due to

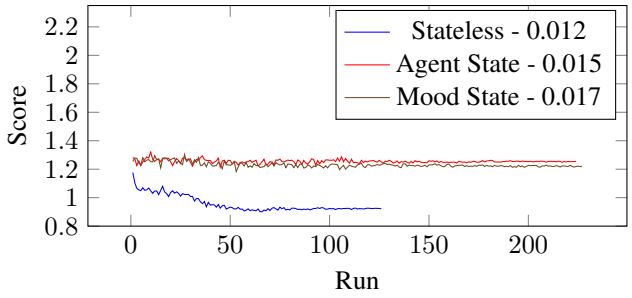

Figure 5: Average scores for Sarsa in the Stag Hunt over each run, with 99% confidence interval of final average score.

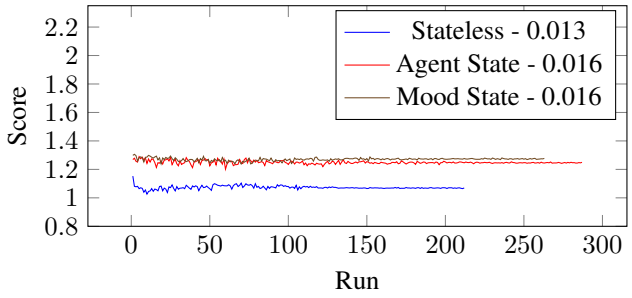

Figure 6: Average scores for memory only in the Stag Hunt over each run, with 99% confidence interval of final average score.

the agents receiving some payoff, the chances of an agent inadvertently choosing cooperation increases. This increase also causes more mutual cooperation raising the moods of those agents even more making it even more unlikely that defection will be chosen. This shows that our mood agents are able to break continued mutual defection and can replace this with sustained mutual cooperation over time.

Conclusion

We have provided a novel adaptation of a classic reinforcement learning algorithm by incorporating within it a model of mood. We have evaluated this extensively in an experimental setting using the Prisoner's Dilemma and Stag Hunt scenarios. We then compared our results to those produced by a Sarsa implementation to investigate whether there was any improvement. We used a number of scenarios that varied the amount of information that was available to an agent in addition to varying the strength of the mood model. We measured improvement in two different ways, namely proportion of cooperation and average reward received by an agent per interaction. In contrast to previous work we have investigated scenarios which have allow agents to be mobile, introducing uncertainty in the interactions. Our study is of use to designers of agent societies, by showing how mobility and mood affect different strategies of the agents.

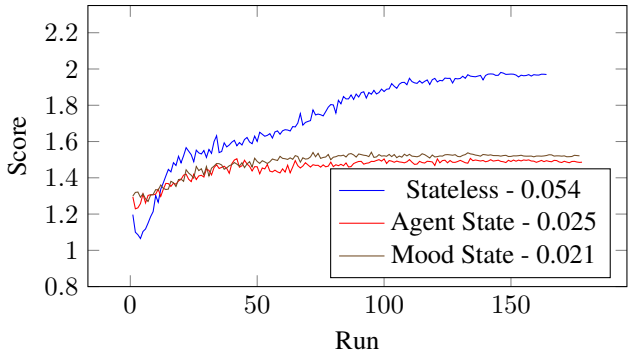

Figure 7: Average scores for mood in the Stag Hunt over each run, with 99% confidence interval of final average score.

By incorporating mood in reinforcement learning we increased the level of cooperation when compared to Sarsa, decreased mutual defection, and introduced more non-mutual actions. We also compared the averages payoffs showing that using mood increased the average payoff when compared to Sarsa. There were differences between the two social dilemmas in regards to how effective the mood model was. Higher levels of cooperation where shown in the Stag Hunt when compared to the Prisoner's Dilemma, due to the payoffs of the game. Additionally we noted the reduction in cooperation Sarsa has when compared to prior research, this is mainly due to the mobility aspect we capture.

The memory aspect can be further developed with additional aspects of psychology being incorporated, such as how recall of past events is affected by the mood (Bower, 1981). The main aim of further research into this area of psychology-driven reinforcement learning would be to reduce the number of non mutual actions so as to give us a more definitive outcome of what the agent has learned. Furthermore we can supplement this research with a mathematical study of the model to allow us to show whether the model is applicable to more social dilemmas. We can extend this further by incorporating other strategies into our scenarios and seeing how our model copes with the addition of these strategies as opponents.

References

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

Bloembergen, D., Ranjbar-Sahraei, B., Bou Ammar, H., Tuyls, K., and Weiss, G. (2014). Influencing social networks: An optimal control study. In *Proc of ECAI'14*, pages 105–110.

Bolton, G. E., Feldhaus, C., and Ockenfels, A. (2016). Social interaction promotes risk taking in a stag hunt game. *German Economic Review*, 17(3):409–423.

Bower, G. H. (1981). Mood and memory. *American psychologist*, 36(2):129.

Collenette, J., Atkinson, K., Bloembergen, D., and Tuyls, K. (2016a). The effect of mobility and emotion on interactions in multi-agent systems. In *Proc of STAIRS'16*.

Collenette, J., Atkinson, K., Bloembergen, D., and Tuyls, K. (2016b). Modelling mood in co-operative emotional agents. In *Proc of DARS'16*.

Crandall, J. W. and Goodrich, M. A. (2005). Learning to compete, compromise, and cooperate in repeated general-sum games. In *Proc of ICML'05*, pages 161–168. ACM.

Devlin, S., Grześ, M., and Kudenko, D. (2011). Multi-agent, reward shaping for robocup keepaway. In *Proc of AAMAS'10*, pages 1227–1228. IFAAMAS.

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

Gray, E. K., Watson, D., Payne, R., and Cooper, C. (2001). Emotion, mood, and temperament: Similarities, differences, and a synthesis. *Emotions at work: Theory, research and applications for management*, pages 21–43.

Haley, W. E. and Strickland, B. R. (1986). Interpersonal betrayal and cooperation: Effects on self-evaluation in depression. *Journal of Personality and Social Psychology*, 50(2):386.

Hepburn, L. and Eysenck, M. W. (1989). Personality, average mood and mood variability. *Personality and Individual Differences*, 10(9):975–983.

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 & emotion*, 14(4):441–472.

Keltner, D. and Gross, J. J. (1999). Functional accounts of emotions. *Cognition & Emotion*, 13(5):467–480.

Leahy, R. L. (2005). Clinical implications in the treatment of mania: Reducing risk behavior in manic patients. *Cognitive and Behavioral Practice*, 12(1):89 – 98.

Levenson, R. W. (1994). Human emotion: A functional view. *The nature of emotion: Fundamental questions*, 1:123–126.

Lloyd-Kelly, M., Atkinson, K., and Bench-Capon, T. (2014). Fostering co-operative behaviour through social intervention. In *Proc of SIMULTECH'14*, pages 578–585. IEEE.

Masuda, N. and Ohtsuki, H. (2009). A theoretical analysis of temporal difference learning in the iterated prisoners dilemma game. *Bulletin of mathematical biology*, 71(8):1818–1850.

Ng, A. Y., Harada, D., and Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In *Proc of ICML'99*, pages 278–287. Morgan Kaufmann.

Ojha, S. and Williams, M.-A. (2016). Ethically-guided emotional responses for social robots: Should i be angry? In *Proc of ICSR'16*, pages 233–242. Springer.

Randløv, J. and Alstrøm, P. (1998). Learning to drive a bicycle using reinforcement learning and shaping. In *ICML*, volume 98, pages 463–471. Citeseer.

Ranjbar-Sahraei, B., Bou Ammar, H., Bloembergen, D., Tuyls, K., and Weiss, G. (2014). Evolution of cooperation in arbitrary complex networks. In *Proc of AAMAS'14*, pages 677–684.

Rinck, M., Glowalla, U., and Schneider, K. (1992). Mood-congruent and mood-incongruent learning. *Memory & cognition*, 20(1):29–39.

Santos, F. C., Santos, M. D., and Pacheco, J. M. (2008). Social diversity promotes the emergence of cooperation in public goods games. *Nature*, 454(7201):213–216.

Santos, R., Marreiros, G., Ramos, C., Neves, J., and Bulas-Cruz, J. (2009). Personality, emotion and mood simulation in decision making. In *Proc of EPIA'09*.

Schwarz, N. (2000). Emotion, cognition, and decision making. *Cognition and Emotion*, 14(4):433–440.

Shteingart, H., Neiman, T., and Loewenstein, Y. (2013). The role of first impression in operant learning. *Journal of Experimental Psychology: General*, 142(2):476.

Skyrms, B. (2004). *The stag hunt and the evolution of social structure*. Cambridge University Press.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT press Cambridge.

Vassiliades, V., Cleanthous, A., and Christodoulou, C. (2011). Multiagent reinforcement learning: Spiking and nonspiking agents in the iterated prisoner's dilemma. *IEEE transactions on neural networks*, 22(4):639–653.

Vaughan, R. (2008). Massively multiple robot simulations in stage. *Swarm Intelligence*, 2(2-4):189–208.

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(12):3083–3096.

Carole Knibbe et al, eds., Proceedings of the ECAL 2017, Lyon, France, 4-8 September 2017,
(Cambridge, MA: The MIT Press, ©2017 Massachusetts Institute of Technology).