EMOTION AS AN ENABLER OF CO-OPERATION

Martyn Lloyd-Kelly[1], Katie Atkinson[1] and Trevor Bench-Capon[1]
[1]*Department of Computer Science, The University of Liverpool, Ashton Street, Liverpool, United Kingdom*
{*mlk5060, katie, tbc*}*@liv.ac.uk*

Keywords: Emotion: Multi-Agent Systems: Agent Models and Architectures: Simulation.

Abstract: Human reasoning and behaviour is undoubtedly influenced by emotions. However, the role of emotion in reasoning has, until recently, been viewed as secondary, with preference given to game theory principles in order to explain how the reasoning of an individual affects sociable interaction and the phenomenon of co-operation. Despite this, development of emotional agent architectures has gained increased interest, resulting in multi-agent systems whose individuals use emotion to aid reasoning and behaviour selection. This paper details a novel emotional agent capable of simple, natural emotional responses to information received from the environment it is situated in. Such an agent is contrasted with the concept of a so-called rational agent, whose reasoning is determined by rational processes that are based upon game theoretic notions. We present a novel test-bed entitled *Tileworld Dilemma* which is inspired by the *Tileworld* test-bed and Robert Axelrod's take on the *Prisoner's Dilemma. Tileworld Dilemma* allows us to research two questions: firstly, is the rational behaviour demonstrated by the most successful strategy in Axelrod's tournament, the "tit-for-tat" strategy, capable of being replicated using simple emotional responses produced by our emotional agent? Secondly, how can emotions enable and promote co-operation between agents in a society? To investigate these questions we pit an emotional agent with a range of emotional characters against the most notable strategies described in Axelrod's tournament and analyse the behaviours and scores demonstrated/obtained by both the individuals and the total system. As a result, we discover that tolerance and responsiveness are integral emotional features with regards to the scoring of agents endowed with functional interpretations of emotions.

1 INTRODUCTION

The interplay and differences between rationality and emotion has long been the subject of philosophical and psychological debates. Current thinking favours the idea that the "gut" (the place in the body where emotions reign supreme) provides human-beings with quick, initial responses to various situations whilst the "head" (the realm of reason) steps in to rationally correct these gut decisions (Gardner, 2008). In many cases however, head does not have a chance to intervene in gut's reasoning, resulting in decisions and actions that can fly in the face of logic and reason. Such behaviour appears to be the product of a number of emotional rules that were created and refined through natural selection during our genesis as a hunter-gatherer society (Tversky and Kahneman, 1983), (Rothman and Schwarz, 1998). Furthermore, even if we take the view that emotions could be important in enabling self-interested members of a society to co-operate, then Robert Axelrod's seminal work *The Evolution of Co-operation* (Axelrod, 1984), shows us that rationality can also achieve this even if an agent is entirely self-interested. The question then is, why consider the use of emotional response in agent systems at all, if rationality is the reasoning engine of choice for humanity?

As argued in (Gardner, 2008) and (Keltner and Gross, 1999), for all their perceived shortcomings, emotional responses must be useful in some way. Indeed, as previously mentioned, emotions appear to play a significant role in our social choices and behaviours; (Fessler and Haley, 2003) show that emotions are key determinants of behaviour in co-operative relationships and (Frank, 1988) asserts that natural selection may have favoured those whose behaviour is determined by their emotions. Herein lies our motivations for including emotions into the decision-making procedures of agents and we take the view that emotions can play a functional role in explaining human behaviour, following (Keltner and Gross, 1999), (Frijda, 1987).

Our contribution is twofold. Firstly, we model emotions so that they are able to play a functional and beneficial role in determining how to respond to information received from the environment and show

that the emergent property of co-operation, achievable by rational behaviour, can also be achieved through emotional responses. Secondly, we propose that emotions are an important contributing factor in enabling the phenomenon of co-operation to occur and that the presence of an emotional agent in a system can be beneficial to the system as a whole. To investigate and make these contributions concrete we describe a novel, implemented test-bed inspired by (Axelrod, 1984) and the *Tileworld* test-bed (Pollack and Ringuette, 1990).

The paper is structured as follows: section 2 outlines our concept of an emotion, the current state of the art in emotional multi-agent systems, the emotional model we propose to use and a brief outline of the inspiration and tools used to create the *Tileworld Dilemma* test-bed. Section 3 describes our test-bed, the research questions this enables us to investigate and the experiment set-ups in detail. Section 4 discusses the experiments, the results obtained and their implications for multi-agent systems. Finally, section 5 summarises the paper and contains a brief outline of our current work.

2 BACKGROUND

The concept of emotion is hotly contested and no real consensual definition has been found. (Wechsler, 1925), and (Kleinginna and Kleinginna, 1981) demonstrate the difficulty in even reaching a consensual definition of the word itself, as a variety of terms exist for the multiple facets of emotion-related terminology. With respect to what an emotion is, (James, 1884) views emotions as being purely physiological i.e. "fear" is a lexical token assigned to describe a collection of bodily reactions such as quickened heart rate, profuse sweating, weakness of limbs etc. produced in response to external stimuli. Other psychologists take the view that emotions act functionally to influence behaviour. (Baumeister et al., 2009) shows that emotions can do this in two ways: emotions can be motivations for behaviour or emotions can be used as an evaluative tool used to explain behaviour that is exhibited in certain situations. Like (Keltner and Gross, 1999) and (Frijda, 1987), we adopt the functional view of emotions and therefore have used this approach when modelling our emotional agent. Whilst we agree that physiological factors are important in any comprehensive account, we do not consider this aspect here.

Emotions have been incorporated into agent architectures by Steunebrink et al. resulting in several papers: (Dastani and Meyer, 2006), (Steunebrink et al., 2007), (Steunebrink et al., 2008), (Steunebrink et al., 2009a) and (Steunebrink et al., 2010). These papers consider small subsets of the 22 emotions defined in the Ortony, Clore and Collins model, otherwise known as the "OCC model" (Ortony et al., 1988). We adopt the idea of implementing small subsets of emotions from this much larger set as it allows a concentration of effort with respect to the emotions chosen, faithful modelling of these emotions and recognition that different emotions may have different functional roles. The OCC model is a framework of emotions which serves as one of the standard psychological frameworks that is adapted for use in computer science. In addition to the work cited above, the model is used by others in the emotional agent systems field including (Nawwab et al., 2010), (Bates, 1994) and (Nissan, 2009). The model is especially attractive to computer scientists as it provides a tractable model of emotions which can be readily adapted for use in the field of artificial intelligence. The model proposes that emotions are a valenced reaction to the consequences of events, the actions of the agent itself, actions of other agents and aspects of objects. From this, we can see that the terminology of the OCC model runs parallel to the terminology used in the field of agent systems (the notion of event consequences can be seen as equivalent to the notion of "goals" with the OCC model recognising the difference between achievement and maintenance goals).

Steunebrink, Dastani and Meyer augment the OCC model by giving a logical formalisation and adapting it where necessary (Steunebrink et al., 2009b). They also adopt a functional view of emotions and prescribe actions that follow after an emotion has been elicited. For example, if fear is elicited with respect to an agent's desirable goal, re-planning may be triggered. The applicability of the OCC from a computer-science standpoint and its extensive use by others is the primary reason that we chose to use the model as the basis for the emotion theory underpinning our emotional agent.

Other agent systems that have attempted to incorporate emotions include (Velásquez, 1997) and (Velásquez, 1998) who classifies six families of emotions which are used in order to bias action-selection in agents using emotional memories. (Oliveira, 2009) also uses emotions in a similar way to facilitate the learning of behaviour rules. The aforementioned papers both cite Damasio, (Damasio, 2005), as support for their decision to use emotions as mediators of behaviour. The work presents Damasio's *somatic marker hypothesis*; the idea that emotions bias our decision-making and behaviour. The somatic marker hypothesis and its usage by others, adds considerable

validity to the extensions of the OCC model proposed by Steunebrink, Dastani and Meyer and provides us with support for using emotions in the way we do. For an extensive recent survey of emotion incorporation in agent systems see (Rumbell et al., 2011).

The work we present here furthers the functional use of emotions as behavioural mediators and uses Axelrod's tournament as a basis to test the effectiveness of using emotions functionally as behavioural modifiers and as an enabler of co-operation. Axelrod played pitted rational strategies against each other in a number of *Iterated Prisoner's Dilemma* games (Poundstone, 1993) to determine which is the most successful. Axelrod's measure of success was to record both a player's individual score and the magnitude of co-operation elicited from the two players (measured by observing the total score of the system). The higher the individual and total system score, the more successful the strategy is. The issue here is that a strategy may be successful against some strategies and unsuccessful against others. The tournament led to the identification of the tit-for-tat strategy as the most successful overall and Axelrod identified four general rules that should be adhered to in order to create a successful strategy. One of these rules states that strategies should not be overly complex; in some cases, strategies were so complex that they might as well have been acting randomly.

This rule provides a basis for us to distinguish between emotional and rational agents. As stated, Axelrod's tournament is populated by agents that rationally condition their behaviour by taking into account their past, present and future payoffs. Therefore, a rational agent determines its behaviour on the basis of payoffs; in contrast, our emotional agent makes no use of the concept of payoffs, they are simply reactive agents inspired by the notions outlined in (Brooks, 1991a) and (Brooks, 1991b). Essentially, the emotional agent's behaviour is a product of its character and its current emotional state (which is determined by its past experience and character) with the layer of rationality associated with consideration of payoffs and history stripped away. The concept of a *current emotional state* is what motivates the design of a novel agent architecture as (Velásquez, 1997; **?**; **?**) all implement an emotional state that is non-persistent. In other words, the emotional state of the agent is confined to particular episodes in the agent's history or present state, an agent does not have a continually updated, general emotional state from which it may choose a particular action.

Whilst Axelrod shows that rational, self-interested behaviour can indeed enable co-operation, (Frank, 1988) argues that such behaviour can be self-defeating and that individuals endowed with emotions are much more likely to establish and maintain co-operation. The question of which emotions contribute to this is debatable but, it is thought that gratitude and anger are important. (Berg et al., 1995) illustrates that financial loss to an individual can be tolerated if the individual's co-operative behaviour is rewarded by way of gratitude. (Fehr and Gächter, 2002) shows that altruistic punishment (whereby an agent suffers loss to enforce a social norm), resulting from anger, is essential in order for human co-operation to flourish. Consequently, we have chosen to focus our efforts on implementing these two emotions and we investigate how *gratitude* and *anger* can influence the total score of the system.

In order to achieve this we required a test-bed to explore these questions. Taking inspiration from (Jiang et al., 2007), we decided to use a Netlogo (Wilensky, 1999) implementation of the *Tileworld* test-bed (Pollack and Ringuette, 1990) and adapt the test-bed context to our particular needs. The concepts of *Tileworld* were used as the game context provides motivation for agents to decide whether to co-operate or defect and it also allows us to test the meta-level reasoning undertaken by agents. The next section details this test-bed, the agent architecture used, and poses the questions of whether rationality can be replicated by emotional response, whether emotional agents can become more successful than the "tit-for-tat" agent with respect to total system score and which character is the most successful. Subsequently, we describe the experiments constructed and run to provide answers to these research questions.

3 TEST-BED IMPLEMENTATION

Our *Tileworld Dilemma* test-bed is defined as a multi-agent game implemented using the Netlogo programming language. Within this game, two agents are situated on a grid which includes two artefacts: one gap and one tile, unlike the standard *Tileworld*, the environment contains no obstacles. When a game is set-up, agents are back-to-back and vision is restricted to 180° in front of each agent. A tile and gap are then generated in each half of the environment so that the location of each is only known to one agent. Figure 1 illustrates the initial set-up of a Tileworld Dilemma game.

When the game begins, each agent generates a percept containing the artefact's type (tile or gap), ID and x/y coordinate. An agent then sends an ask locution to its opponent and a response is provided by way of an inform locution. Communication is facili-

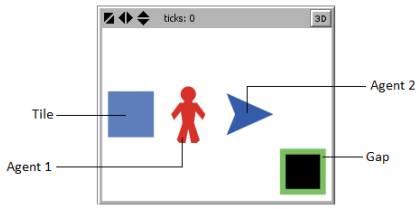

Figure 1: Tileworld Dilemma initial set-up.

tated by the use of a FIPA ACL-like message passing system for Netlogo (Sakellariou, 2010). Typical ask and inform locutions take the following forms:

```
[ask sender:3 content: 0 2 receiver:2]
[inform sender:2 content: 0 0 2 receiver:4]
```

When sending an ask locution, the content field contains an x/y-coordinate whereas the content field of an inform locution is composed of the artefact's ID number and x/y-coordinate. An inform locution's content field can contain true information, which we call *co-operation*, or false information, which we call *defection*; terms taken from the *Prisoner's Dilemma*. The decision to co-operate or defect is dependent upon the agent's current strategy (see table 3 for descriptions of strategies) or emotional state. Six strategies are direct replicas of the most notable submissions detailed in Axelrod's tournament with the addition of our emotional strategy/agent. The emotional agent is capable of different behaviour depending upon its *character*, which determines how quick or slow an agent is to reward or punish. We say that the slower an agent is to punish defection, the more *tolerant* it is, and the quicker it is cooperate the more *responsive* it is. Thus, if an emotional agent is of character 1 - our least tolerant and most responsive agent - only one defection is required for its anger to become sufficient to change its behaviour to defection. Similarly, an opponent only needs to co-operate once in order to elicit sufficient gratitude to cause the agent to co-operate. Conversely, an emotional agent of character 9 - our most tolerant and least responsive character - requires three co-operations/defections before the agent feels sufficient gratitude or anger to change its behaviour. Details of all nine characters are given in table 1 for reference. Each emotional agent also has an initial disposition which can be set to co-operate or defect; this value is used to determine the behaviour on the initial round of every game.

The receiver field contained within the inform locution contains the ID number of a third disembodied agent: the program mediator. The mediator, unlike the two agents, is able to see the whole environment and can therefore assess and report the validity of responses to the players. The mediator is also responsible for distributing payoffs; these payoffs are used to

Table 1: Emotional agent character (ch.#) descriptions.

		If defecting, #co-ops required to co-op.		
		1	**2**	**3**
If co-op, #defects required to defect.	**1**	Ch.1	Ch.2	Ch.3
	2	Ch.4	Ch.5	Ch.6
	3	Ch.7	Ch.8	Ch.9

compare the strategies employed by agents in exactly the same way as in Axelrod's tournament by the rational agents, while the emotional agents react to the nature of the action rather than the payoff. Payoffs are distributed exactly as for the *Prisoner's Dilemma* (see table 2 for details).

Table 2: Prisoner's Dilemma payoff matrix.

	Co-operate$_j$	**Defect**$_j$
Co-operate$_i$	$3_i, 3_j$	$0_i, 5_j$
Defect$_i$	$5_i, 0_j$	$1_i, 1_j$

A typical round in the *Tileworld Dilemma* progresses as follows:

- Agents, artefacts and mediator are placed in the environment.
- Agents generate percepts regarding artefacts and ask locutions are sent to opponents.
- Agents consult strategy and respond by sending inform locutions to the program mediator.
- Mediator determines validity of inform locutions, distributes payoffs and informs agents of the opponent's response validity.

The number of rounds and games played by the agents is not known/accessible to either agent/mediator in order to negate the "backward induction paradox" (Petit and Sugden, 1989), a relevant issue when using rational agents to play the *Iterated Prisoner's Dilemma*.

3.1 Emotional Agent Description

As explained in section 2, the emotional agent present in the *Tileworld Dilemma* is inspired by the reactive agent architecture proposed by Brooks in (Brooks, 1991a) and (Brooks, 1991b). The agent reacts to both visual input, producing percept generation, and textual input, which may produce an emotional response that has the potential to cause a change in behaviour (dependent upon the emotional agent's character).

Table 3: Descriptions of strategies present in the Tileworld Dilemma.

Strategy	Behaviour Description
Emotional	Dependent upon character selected.
Mendacious	Always lies.
Veracious	Always tells the truth.
Random	Has a 1 in 2 chance of defecting/co-operating in each round.
Tit-for-tat	Co-operates on the first round, n. In next round it mimics the opponent's behaviour from the previous round.
Tester	Defects on first round n, if the opponent co-operated on round n then the agent co-operates on rounds n + 1 and n + 2 and defects on round n + 3. If, the opponent defects on round n, the agent plays tit-for-tat for the rest of the game.
Joss	Plays tit-for-tat, however, has a 1 in 10 chance of defecting on a round.

Emotions may only be elicited by one input: the validity of the inform locution sent by the agent's opponent. Each emotional character has varying degrees of sensitivity to co-operation and defection i.e. character 1 is the most sensitive to both whilst 9 is the least sensitive. The sensitivity of an agent towards various emotion eliciting factors of a situation dictates how easy it is for an emotional threshold to be reached. This idea comes from the OCC model where it is proposed that emotions have the potential to be elicited in any situation, but only if an emotional threshold is reached. As explained in section 2, we have endowed the implemented emotional agents with two of the twenty-two emotions proposed by Ortony et al.: gratitude and anger. The OCC model defines these emotions as:

- Gratitude: approving of someone else's praiseworthy action and being pleased about the related desirable event.
- Anger: disapproving of someone else's blameworthy action and being displeased about the related undesirable event.

Gratitude can only be elicited by receiving co-operation whereas anger can only be elicited in response to defection; we model these emotions as having opposite functionality. For example, if gratitude is elicited and the agent is currently defecting the agent will co-operate i.e. its function is to reward others for their co-operation. Conversely, if anger is elicited and the agent is currently co-operating the agent will defect i.e. its function is to punish others for their defection. Therefore, the emotions have a functional influence on the behaviour of the agent, and the current behaviour of the agent is a product of past emotional experience and its character. Thus, the emotional agent does not react according to its current or expected score, but to the validity of the information sent to it by its opponent. This removes the layer of rationality present in rational agents where payoffs are considered and strategies are adopted accordingly.

3.2 Experiment Set-up

In order to provide results for the research questions outlined in section 3.3 each rational agent will play against all other rational agents for 5 games of 200 rounds (as in Axelrod's tournament). Then, an initially co-operative emotional agent with character 1 will play against each rational agent for the same number of games/rounds. The emotional agent's character is then increased by 1 and all agents are played again until character 9. It should made clear that when an agent's character is increased by 1, its current emotional state is erased so that its new emotional state is not affected by events that occurred in the preceeding game. The emotional agent's initial disposition will then be changed so that it defects initially and its character reset to 1. All rational agents will then be played against all emotional agent characters again.

It should be noted that random and joss agents have their random seed numbers set when they are created. This ensures that the behaviour of such agents is consistent between experiments as the number provided to the random seed produces the same order of outputs (behaviour) whenever the random number generator is used. This feature is desirable as, although we need random behaviour, we need this behaviour to be consistent across particular experiments for comparative purposes.

3.3 Research Questions

Within our general hypothesis we have identified two specific research questions that the implemented testbed aims to answer. These questions are presented below along with information that we will extract from the system which we deem to be relevant to each question.

1. Is there an emotional agent character capable of replicating the behaviour of an agent which uses

the most successful strategy in Axelrod's tournament: the tit-for-tat strategy, and if so, how is this achieved?

2. Are there any other emotional characters which improve upon the success of the tit-for-tat strategy with respect to the total system score, and if so, why?

With respect to the first question we believe that we can ascertain if rational behaviour is being replicated by simply analysing the individual scores and total system scores for the tit-tat agent and each emotional agent set-up after five games have been played. If the scores of an emotional agent set-up are observationally equivalent to a tit-for-tat agent we will attempt to explain why by comparing the salient aspects of both the emotional agent's character and the tit-for-tat agent's strategy. The second question can be initially answered by taking an average of the total system scores for each of the five games that the agent plays in. These average total system scores for each rational agent that the emotional agent plays against are then summed together to produce the *aggregated average total system score*. If an emotional agent's character provides a greater aggregated average total system score than that achieved by the tit-for-tat agent then this character will be flagged and further in-depth analysis of relevant data will be conducted to answer the remainder of the question.

4 RESULTS AND DISCUSSION

With respect to the question of whether the tit-for-tat strategy can be replicated by emotional response, the results obtained from the *Tileworld Dilemma* are clear-cut. An initially co-operative emotional agent with character 1 replicates the behaviour of the rational tit-for-tat agent exactly. To demonstrate this, we present tables 4 and 5 which contain the average individual scores of the initially co-operative emotional agent with character 1 and the tit-for-tat agent versus the random and joss agents respectively. We have chosen to only present the results from playing these two agents as the behaviour of the random and joss agents is non-deterministic whilst the behaviour of every other agent in the simulation is completely deterministic. Therefore, playing against these two agent types gives the greatest potential for disparateness to exist between the scores of the emotional agent and tit-for-tat agent. Consequently, by presenting these two graphs as evidence we can assert that the behaviour of the tit-for-tat agent is exactly replicated by the emotional agent with the set-up described.

Table 4: Individual scores of intially co-operative emotional agent with character 1/tit-for-tat agent vs. random agent.

	Game Number				
Agent	**1**	**2**	**3**	**4**	**5**
Emotional	462	466	448	445	424
Tit-for-tat	462	466	448	445	424

Table 5: Individual scores of intially co-operative emotional agent with character 1/tit-for-tat agent vs. joss agent.

	Game Number				
Agent	**1**	**2**	**3**	**4**	**5**
Emotional	219	213	213	255	242
Tit-for-tat	219	213	213	255	242

As can be seen in tables 4 and 5, the scores of the emotional agent and tit-for-tat agent exactly overlap showing that their behaviour is undisputedly the same. Explanation of these results is elementary: whereas the tit-for-tat agent responds to its payoffs, the emotional agent responds to information sent to it (as detailed in section 3.1). Therefore, both agents react in exactly the same way to inputs that are of different types but which will arise from the same situations. To clarify, if the tit-for-tat agent observes that it has scored 0 in a round when it is currently co-operating or 1 if is it defecting, then it can safely infer that the opponent is defecting therefore its behaviour will switch to defection. Similarly, if the program mediator informs an emotional agent with character 1 that the opponent has defected, then the emotional agent will defect immediately in the next round.

We now address the question of whether any other emotional character set-up is more successful with respect to maximising the total system score when playing against periodically defecting strategies than the set-up previously discussed. To determine this, we measure success in terms of total system payoff or, more specifically, the aggregated average total system score (the sum of each average total system score achieved by an agent). As demonstrated in table 6, we find that an initially co-operative agent with character 7 - the most tolerant and most responsive - offers the greatest aggregated total average system score so a more successful strategy does indeed exist. To explain this outcome we have identified three criteria which must be considered and discussed in turn: fairness, readiness to co-operate and tolerance.

We define *fairness* as the extent to which all members of a system are equal; in the context of the *Tileworld Dilemma*, the fairest system possible is one where each agent has an equal score at the end of each game. Systems that are maximally fair are achieved by agents who employ strategies that are quick to punish and defect (as noted by Axelrod). If such a strat-

Table 6: Intially co-operative emotional agent aggregated average total system scores.

Character	Aggregated Average Total System Score
1	5230.80
2	5069.80
3	4979.80
4	5774.80
5	5241.80
6	5140.80
7	5895.60
8	5328.80
9	5235.80

egy is used by both players and a cycle of defection is locked into on the first round, then each player's score at the end of a game will be 200. Whilst this is individually fair, the final system score is relatively low. For a player who wishes to achieve system fairness and maximise the score of each player then the best possible score that can be achieved is 600, which is achieved by players immediately locking into a co-operation cycle on the first round and maintaining this for a full game. We observe that the only agent pairs to do this are those that co-operate initially, those that are quick to punish/defect and those that always co-operate, no matter what i.e.:

- Initially co-operative emotional agent with any character and tit-for-tat agent.
- Initially co-operative emotional agent with any character/tit-for-tat agent and veracious agent.

However, as mentioned above, such behaviour does not maximise the system's score when agents that seek an advantage, such as the random, tester and joss agents, are also present (see table 6); from the system's view, achievement of a good system score requires two goals to be achieved:

- Co-operation must be established between the members of the system.
- Co-operation must be maintained between the members of the system.

The score of a system is increased if agents lock into cycles of co-operation quickly and break them slowly. Therefore, readiness to co-operate and tolerance of defection are *both* important factors. If we compare the average total scores for an initially co-operative emotional agent of character 7 to an initially lying emotional agent of character 7 (see table 7) then the effect of being quick to co-operate becomes clear.

If an agent initially defects, co-operation cycle establishment is delayed, resulting in lower total system scores as it becomes more likely that the players will establish cycles of defection. Conversely, the quicker an agent is to co-operate and forgive its opponent, the quicker a co-operation cycle is established. Therefore, by co-operating initially an agent is more likely to find concurrent co-operation in a round and establish a co-operation cycle early in the game (important as the number of rounds in a game is finite); table 8 clearly illustrates this point. The same pattern also holds true for initially defective/co-operative emotional agents with characters 1-3/4-6.

It is not enough to simply establish a cycle of co-operation; in order to maximise the score of the system then the established co-operation cycle must be maintained, even when the other player temporarily defects (as self interested agents will tend to do). If we consider the scores of emotional agents with characters 1, 4 and 7 displayed in table 6, we observe that as an agent becomes more tolerant to defections, the greater the aggregated average total system score becomes. If we then consider the individual scores which are aggregated together for the initially co-operative emotional agent of character 7 (see table 9) we can see that character 7 sacrifices system fairness by taking a reduced score in order to maximise the total system score. This phenomenon of *tolerance* is the crucial difference between character 7 and characters 1 and 4. Therefore, we can see that increased levels of tolerance are integral to maximising the total system score, if playing against agents that periodically defect.

By being tolerant an agent enables the maintenance of a co-operation cycle. Whilst the fairest system possible entails the deployment of a strategy that is quick to reward and quick to punish, such behaviour breaks co-operation cycles quickly causing lower total system scores to be achieved. By one agent continuing to co-operate in the face of defection the system scores five rather than two, so that when the defector decides to co-operate again and it is met with co-operation, a total system score of six is achieved. A drawback to becoming more tolerant however is suffering a reduction in the tolerant agent's individual score; table 10 illustrates the extent to which this occurs.

Table 10 offers some interesting results, especially if we consider those scores that pertain to the emotional agent playing against the random agent. We observe that the average individual score of each agent decreases as tolerance to defection increases yet, as tolerance is increased the rate at which the average individual score decreases slows; this can also be observed in figure 2 and table 11. The salient point here is: when the opponent is not a veracious or tit-for-tat

Table 7: Comparison of the average total scores of an initially co-operative emotional agent of character 7 and an initially defective emotional agent of character 7.

	Opponent					
Ini Dis.	**Mendacious**	**Veracious**	**Random**	**Tit-for-tat**	**Tester**	**Joss**
Co-op	409	1200	1002.8	1200	1111	972.8
Defect	400	1199	1001.8	1198	400	968.6

Table 8: Comparison of the average total scores for intially co-operative emotional agents with characters 7, 8 and 9.

	Opponent					
Character	**Mendacious**	**Veracious**	**Random**	**Tit-for-tat**	**Tester**	**Joss**
7	409	1200	1002.8	1200	1111	972.8
8	409	1200	942	1200	1089	488.8
9	409	1200	902	1200	1036	488.8

Table 9: Average Individual scores of initially co-operative emotional agents with characters 1 and 7.

	Opponent$_j$					
Character$_i$	**Mendacious**	**Veracious**	**Random**	**Tit-for-tat**	**Tester**	**Joss**
1	199_i, 204_j	600_i, 600_j	449_i, 451_j	600_i, 600_j	533_i, 533_j	228.4_i, 233.4_j
7	197_i, 212_j	600_i, 600_j	372.4_i, 630.4_j	600_i, 600_j	443_i, 668_j	449.4_i, 523.4_j

agent, there is a trade-off between fairness and total system score. From table 10 we can calculate this trade-off exactly: for every point earned by the system, the emotional agent must lose two points from its individual score. This raises the question: how much of a reduction in fairness is acceptable to achieve these system gains?

Table 10: Average individual score of initially co-operative emotional agents with character 1, 4 and 7 when played against random, tester and joss agents.

	Opponent		
Character	**Random**	**Tester**	**Joss**
1	449	533	228.4
4	398.2	465	417.2
7	372.4	443	449.4

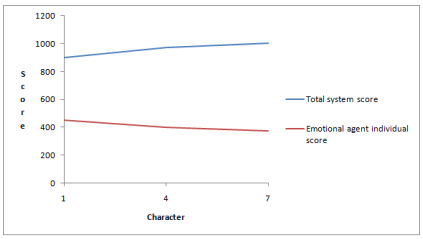

Figure 2: Total system score achieved when initially co-operative emotional agents of characters 1, 4 and 7 plays against a random agent plotted against the individual score of the initially co-operative emotional agents.

It is worth mentioning that the situation is different when the initially co-operative emotional agents with characters 1, 4 and 7 play against a joss agent. As the emotional agents become more tolerant, the emotional agent's average individual score increases

Table 11: Percentage of total system score owned by the initially co-operative emotional agents of characters 1, 4 and 7 when playing aginst the random agent.

	%Total Score Owned	
Character	**Emotional**	**Random**
1	49.9	50.1
4	40.9	59.1
7	37.1	62.9

(see figure 3 and table 10). This is due to the joss agent's behaviour, which enables the maintenance of co-operation cycles in the face of rare, one-off, periodic defections.

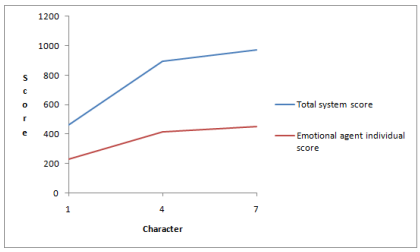

Figure 3: Total system score achieved when initially co-operative emotional agents of characters 1, 4 and 7 plays against a joss agent plotted against the individual score of the initially co-operative emotional agents.

In order to determine when the trade-off between an individual's score and the system's score becomes unacceptable we need to make note of a number of thresholds. To do this we consider a number of various maximal and minimal scores that can be achieved/tolerated for/by each entity in the *Tileworld Dilemma*; table 12 below illustrates these values:

Table 12: The differing threshold values present in the *Tileworld Dilemma* along with how they are derived and their maximum/minimum values.

Threshold Value	How Derived	Max.	Min.
Average Agent 1 Score (A1)	A1 Individual Score	1000	0
Average Agent 2 Score (A2)	A2 Individual Score	1000	0
Average System Score	A1 + A2	1200	400
Average Fairness Score	A1/A2	1	0

The best possible score that an individual agent can achieve is 1000 whilst the worst is 0, achieved when a mendacious strategy is played against a veracious strategy. An individual score of 0 is the worst scenario possible; yet, the lowest acceptable score that can be achieved by a single agent is 200, caused by two players locking into a defection cycle for a whole game. The best possible score from the system's perspective is 1200, achieved when two agents co-operate initially and lock into a co-operation cycle for a whole game and the worst score is achieved by two agents locking into a defection cycle for a whole game, leading to a total system score of 400. The rating of fairness ranges from 0, to 1, the closer to 1 the more equal the two player's scores are.

Therefore, we can say that an initially co-operative emotional agent with character 7 is more successful than an initially co-operative emotional agent with character 1 due to its ability to quickly establish and maintain co-operation. Granted, the total system scores produced are not fairly distributed: against a random agent the system/fairness value for an initially co-operative emotional agent of character 7 is 0.59, whereas for an initially co-operative emotional agent of character 1 the system/fairness value is 0.99 (see table 12 for details on how fairness is calculated. Despite this, the system total achieved by an initially co-operative emotional agent of character 7 is much higher than that achieved by its less tolerant peers. It is conceivable that more tolerant agents will produce greater total system scores at the expense of fairness, but only until a certain point i.e. when their individual score passes below the threshold of 200; after this the trade-off becomes definitely unacceptable since consistent defection produces a better result and there are no individual gains from co-operating.

5 CONCLUSION

Our experiments have demonstrated that the rational behaviour exhibited by the tit-for-tat strategy present in (Axelrod, 1984) can be replicated by an initially co-operative emotional agent with character 1 i.e. an agent with a low anger threshold resulting in immediate punishment in response to defection and a low gratitude threshold resulting in immediate reward in response to co-operation. Furthermore, we have also shown that when playing against strategies that intersperse co-operation with periodic defection a readiness to co-operate and degree of tolerance are key characteristics that are required in order to maximise the total score of the system. However, by becoming increasingly tolerant and remaining just as ready to co-operate, one must expect to suffer a loss with respect to one's individual score. Consequently, such altruism is only demonstrated if it is worthwhile to do so.

Following on from this work, we have implemented and begun testing an extension to the *Tileworld Dilemma* entitled *Emotional Population*. This test-bed consists of a population of agents (338 in total) that are entirely emotional and capable of being initialised with individual characters in exactly the same way as described in this paper. The *Emotional Population* however incorporates into the existing emotion set consisting of anger and gratitude the additional emotion of admiration. Admiration has the potential to be elicited when an agent's neighbour obtains the highest individual score after n number of rounds, but, as with anger and gratitude, agents have varying degrees of sensitivity with respect to admiration. If admiration is elicited then the evaluating agent will change its initial disposition and emotional character to become more like the successful agent. Through this new scenario we aim to analyse which emotional characters become prevalent in a population and how, as well as investigating the conditions and number of initial co-operators/defectors must be present in a population before co-operation/defection becomes the dominant strategy used.

REFERENCES

Axelrod, R. (1984). *The Evolution Of Cooperation*. Basic Books, Inc.

Bates, J. (1994). The role of emotion in believable agents. *Communications of the ACM*, 37(7):122–125.

Baumeister, R. F., DeWall, C. N., Vohs, K. D., and Alquist, J. L. (2009). *Does Emotion Cause Behavior (Apart from Making People Do Stupid, Destructive Things)?*, chapter 7, pages 119–136. Oxford University Press.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity and social history. *Games and Economic Behaviour*, 10:122–142.

Brooks, R. A. (1991a). Intelligence without reason. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence (ICJAI -91)*, pages 569–595. Morgan Kaufmann.

Brooks, R. A. (1991b). Intelligence without representation. *Artificial Intelligence*, 47:139–159.

Damasio, A. (2005). *Descartes' Error: Emotion, Reason and the Human Brain*. Penguin.

Dastani, M. and Meyer, J.-J. C. (2006). Programming agents with emotions. In Brewka, G., Coradeschi, S., Perini, A., and Traverso, P., editors, *European Conference on Artificial Intelligence*, volume 141 of *Frontiers in Artificial Intelligence and Applications*, pages 215–219. IOS Press.

Fehr, E. and Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415:137–140.

Fessler, D. M. and Haley, K. J. (2003). *Genetic and Cultural Evolution of Cooperation*. Cambridge, MA: MIT Press.

Frank, R. H. (1988). *Passions Within Reason: The Strategic Role of the Emotions*. W.W. Norton & Company.

Frijda, N. H. (1987). *The Emotions*. Cambridge University Press.

Gardner, D. (2008). *Risk*. McClelland and Stewart Ltd.

James, W. (1884). What is an emotion? *Mind*, 9:188–205.

Jiang, H., Vidal, J. M., and Huhns, M. N. (2007). EBDI: An architecture for emotional agents. In *Proceedings of the 6th Internationl Joint Conference on Autonomous Agents and Multi-Agent Systems, AAMAS '07*, pages 38–40. ACM Press.

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

Kleinginna, P. R. and Kleinginna, A. M. (1981). A categorized list of emotion definitions with suggestions for a consensual definition. *Motivation and Emotion*, 5(4):345–379.

Nawwab, F. S., Bench-Capon, T., and Dunne, P. E. (2010). Emotions in rational decision making. In *Lecture Notes in Computer Science*, volume 6057, pages 273–291. Springer.

Nissan, E. (2009). Computational models of the emotions: From models of the emotions of the individual to modelling the emerging irrational behaviour of crowds. *AI and Society*, 24(4):403–414.

Oliveira, F. S. (2009). Modeling emotions and reason in agent-based systems. In *Computational Economics Vol. 35*, pages 155–164. Springer.

Ortony, A., Clore, G. L., and Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge University Press.

Petit, P. and Sugden, R. (1989). The backward induction paradox. *The Journal of Philosophy*, 86(4):169–182.

Pollack, M. and Ringuette, M. (1990). Introducing the tileworld: Experimentally evaluating agent architectures. In *In Proceedings of the Eighth National Conference on Artificial Intelligence*, pages 183–189. AAAI Press.

Poundstone, W. (1993). *Prisoner's Dilemma*. Anchor.

Rothman, A. and Schwarz, N. (1998). Constructing perceptions of vulnerability: Personal relevance and the use of experiential information in health judgments. *Personality and Social Psychology Bulletin*, 24(10):241053–1064.

Rumbell, T., Barnden, J., Denham, S., and Wennekers, T. (2011). Emotions in autonomous agents: Comparative analysis of mechanisms and functions. *Autonomous Agents and Multi-Agent Systems*, 23:1–45.

Sakellariou, I. (2010). An attempt to simulate fipa acl message passing in netlogo. `http://users.uom.gr/~iliass/projects/NetLogo/FIPA_ACL_MessagesInNetLogo.pdf`. Date Accessed: 9/10/2010.

Steunebrink, B. R., Dastani, M., and Meyer, J.-J. C. (2007). A logic of emotions for intelligent agents. In *22nd Conference on Artificial Intelligence*, pages 142–147. AAAI Press.

Steunebrink, B. R., Dastani, M., and Meyer, J.-J. C. (2008). A formal model of emotions: Integrating qualitative and quantitative aspects. In Ghallab, M., Spyropoulos, C. D., Fakotakis, N., and Avouris, N. M., editors, *European Conference on Artificial Intelligence*, volume 178 of *Frontiers in Artificial Intelligence and Applications*, pages 256–260. IOS Press.

Steunebrink, B. R., Dastani, M., and Meyer, J.-J. C. (2009a). A formal model of emotion-based action tendency for intelligent agents. In Lopes, L. S., Lau, N., Mariano, P., and Rocha, L. M., editors, *EPIA*, volume 5816 of *Lecture Notes in Computer Science*, pages 174–186. Springer.

Steunebrink, B. R., Dastani, M., and Meyer, J.-J. C. (2009b). The OCC model revisited. In *4th Workshop on Emotion and Computing*. Paderborn.

Steunebrink, B. R., Dastani, M., and Meyer, J.-J. C. (2010). Emotions to control agent deliberation. In van der Hoek, W., Kaminka, G. A., Lespérance, Y., Luck, M., and Sen, S., editors, *9th International Conference on Autonomous Agents and Multiagent Systems*, volume 1, pages 973–980. IFAAMAS.

Tversky, A. and Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90(4):293–315.

Velásquez, J. D. (1997). Modeling emotions and other motivations in synthetic agents. In *Proceedings of the 14th National Conference on Artificial Intelligence*, pages 10–15. AAAI Press.

Velásquez, J. D. (1998). When robots weep: Emotional memories and decision-making. In *Proceedings of the 15th National Conference on Artificial Intelligence*, pages 70–75. AAAI Press.

Wechsler, D. (1925). What consititues an emotion? *Psychological Review*, 32(3):235–240.

Wilensky, U. (1999). Netlogo. `http://ccl.northwestern.edu/netlogo`. Date Accessed: 23/6/2010.