Imperial College London
Department of Computing

Solving Stochastic Games for Child-Parent Interaction

by

Cai Zhou

Submitted in partial fulfilment of the requirements for the MSc Degree in Computing Science/ Computational Management Science of Imperial College London

September 2012

Abstract

Stochastic games, which generalize the classical matrix games between players, have been increasingly employed in multi-agent systems to model learning behaviour.

In this project, we extend Q-learning to multiagent system, based on the framework of stochastic games. This framework provides a way of describing the dynamic interactions of agents in terms of individuals' Markov decision processes and the aim is make both the parent and child are able to learn. Game theoretic models based on the works by David Cittern are all simulating using stochastic games framework. Next, we consider an iterative game environment or how attachment classifications may come to change over repeated interactions, whereby the parent and child's payoff matrix can evolve over time. We show how these mechanisms can lead an initially avoidant or disorganized dyad into a secure attachment style, and consider how this evolution is formed by both immediate and future rewards and the set of outcomes that are encourage which uses reinforcement rules to achieve. Finally, simulation results provide primary results analysis and we observe the change of matrix evolving progress using reinforcement rules for multi-agent in Child-Parent Interaction games.

Acknowledgements

I would like to thank Professor Abbas Edalat for his patient guidance with the project and I would like to thank David Cittern for his parts of code provided the basic infrastructure of my programming work and his project provides my main base of my work. Finally, I would like to thank my family and friends for their support this past year.

Contents

CHAPTER 1

Introduction

Attachment is an emotional bond to another person. Psychologist John Bowlby was the first attachment theorist, describing attachment as a "lasting psychological connectedness between human beings" [7]. Bowlby believed that the earliest bonds formed by children with their caregivers have a tremendous impact that continues throughout life. According to Bowlby, attachment also serves to keep the infant close to the mother, thus improving the child's chances of survival.

In her 1970's research, psychologist Mary Ainsworth expanded greatly upon Bowlby's original work. Her groundbreaking "Strange Situation study revealed the profound effects of attachment on behavior. In the study, researchers observed children between the ages of 12 and 18 months as they responded to a situation in which they were briefly left alone and then reunited with their mothers [6].

Based upon the responses the researchers observed, Ainsworth described three major styles of attachment: secure attachment, ambivalent-insecure attachment and avoidant-insecure attachment. Later, researchers Main and Solomon [10] added a fourth attachment style called disorganized-insecure attachment based upon their own research. A number of studies since that time have supported Ainsworth's attachment styles and have indicated that attachment styles also have an impaction behaviours later in life.

Although internal working models are formed at an early age within the child, research has shown that changes in the behaviour, attitude or environment of the parent can lead to changes in the style of attachment that they have with their child [11]. Since the effects of these early interactions on both individuals and society are great, it is important that we understand with some precision both how these various attachment styles are formed, and how they might come to change.

A further, important attachment classification is disorganisation; a more complicated category in which children have been observed to exhibit bizarre, contradictory and disturbing behaviours as a response to the triggering of their attachment system. Disorganised attachment has been linked to parents with unresolved trauma and borderline personality disorders, and their children are at elevated risk of developing pathological problems [23].

David Cittern captured secure, avoidant and ambivalent attachment to present models for the previously unconsidered, but clinically very important, disorganised attachment style[1].

In this project we will use game theory to understand and model parent-child interactions within an attachment context. In particular, we will base on previous work done on capturing secure, avoidant and ambivalent , disorganised attachment to quantify models for the previously

unimplemented in Cittern's project, avoidant and disorganised attachment style in multiagent stochastic game framework. Furthermore, we will take the first steps in considering the mechanisms by which these attachment classifications can come.

1.1 Report Structure

The chapters of this report are structured as follows:

Background

We begin by giving an overview of attachment theory, including its origins, experimental techniques, classification system and arguments for causality. Then we move on to give an overview of game theory, which is the tool we will use to model attachment interactions, and explain concepts such as matrix game, Nash equilibrium and Pareto optimality. Next we will look in some detail at previous work done on capturing secure, avoidant, and ambivalent and disorganized attachment in a game theoretic model. Finally, for each model we classify the resulting ordinal games according to pure Nash equilibria, determine the mixed strategies, and discuss the attachment styles that these various equilibria represent.

Multiagent Reinforcement Learning

In this chapter, we introduce the stochastic games framework and simple Q-learning algorithms used in multi-agent game following the stochastic games, the properties of Q-learning algorithms, and the probability -choice strategy in the game. In order to expand the possibility of further developments of modeling, we list some classical and some newer algorithms that can be applied in multi-agent context, using the framework of stochastic games.

Iterated Game and Simulation

Next we focus our attention on understanding how these attachment styles may come to change. We present a model whereby the child plays a standard iterated game and the both parent and child are subject to controlled payoff reinforcements, resulting in an evolution of their payoff matrix. In addition we outline a simple learning algorithm to show how we can model the agents' decision making process in terms of the state of their ordinal payoff matrix both in single round and multiple rounds, which used to observe the progress of evolving and how many average rounds leading to convergence of games. We show how the combination of such mechanisms can result in the evolution of the dyad's attachment style from avoidant to secure, and disorganised behavior or combined to secure as well.

Evaluation

There are no use quantitative or qualitative methods for evaluation. We just give a brief assessment by own and show the strengths and weakness in objective view.

Conclusions and Future Work

Finally we give an overall summary of our models and findings in these three main areas, and consider improvements and advances that could be made.

CHAPTER 2

Background

2.1 Attachment Theory

Attachment theory is based on the joint work of J. Bowlby and M.S. Ainsworth. Its development history begins in the 1930s, with Bowlby's growing interest in the link between maternal loss or deprivation and later personality development and with Ainsworth's interest in security theory.
Although Bowlby's and Ainsworth's collaboration began in 1950, it entered its most creative phase much later, after Bowlby had formulated an initial blueprint of attachment theory, drawing on ethology, control system theory, and psychoanalytic thinking, and after Ainsworth conducted the 1st empirical study, known as "Strange Situation Protocol" of infant-mother attachment pattern.
The central theme of attachment theory is that a primary caregiver who is available and responsive to their infant's needs establishes a sense of security. The infant knows that the caregiver is dependable, which creates a secure base for the child to then explore the world. In addition the attachment which the child forms with its primary caregiver in the first year of life will not only affect their relationship with the caregiver, but will affect the attachments they form for the rest of their life. Bowlby also concluded that the biological function of attachment is survival, and the psychological function is to gain security. Attachment to the caregiver keeps the infant close, thus improving their chances of survival, and that it was for this very reason that evolution had equipped the child with the means of attracting the parent's attention (e.g. crying) during times of distress. [9]
Bowlby's work caused a virtual revolution in society's care of children, radically influencing areas such as the formulation of policies on hospital visiting by parents and provision for children's play while in hospital, as well as policies on adoption, fostering and familial intervention by social workers.

2.1.1 The Strange Situation Protocol

Bowlby proposed four stage model for the development of attachment between mother and new-born child [1]: Phase 1 (Pre-attachment); Phase 2 (Attachment-in-the-making); Phase 3 (Clear-cut attachment); Phase 4 (Goal-corrected partnership). Phase 3 (Clear-cut attachment) is a crucial stage: from 7 to 24 months old the child's individual interactions have become organised into patterns, which define lasting relationships. A significant development occurs between 7 to 8 months, where the child develops the capability to miss the mother when they are absent. People

are now no longer interchangeable to the child, and a specific attachment to a particular individual (the caregiver) now exists.

In Ainsworth's "Strange situation" study, researchers also observed children between the age of 12 and 18 months as they responded to a series of high and low anxiety situation, paying particular attention to the child's interaction with its mother and the quality of the child's exploration of a new environment. In order to categorized the behaviour of children and find the causality of the behaviours occurring, Ainsworth devised a procedure known as the Strange Situation Protocol as the laboratory portion of her larger study, to assess separation and reunion behaviour [6]. The Strange Situation is designed to elicit attachment behaviour through exposure to eight 3-minute episodes that are moderately increasing stressful for the infant. The Strange Situation Experiment, as a research tool used to assess attachment patterns in infants and toddlers, is not strictly an experiment but rather it is a standardised laboratory procedure that presents infants with a controlled and replicable set of experiences. The protocol is conducted in following steps:

Episode	**Personae/ Duration time**	**Movements**
1	Mother (or other familiar caregiver), Baby, and Experimenter (30 seconds)	Experimenter introduces mother and baby to experimental room, then leaves
2	Mother, Baby (3 mins)	Mother is nonparticipant while baby explores; If necessary, play is stimulated after 2 min.
3	Mother, Baby, Stranger (3 mins or less)	A stranger enters the room and silent in first minute, converse with mother in second minutes, then attempts to approach the baby; After 3 minutes mother leaves the room unobtrusively
4	Stranger, Baby (3 mins)	The stranger attempts to comfort/engage the child. Then mother comes back
5	Mother, Baby (3 mins)	Mother attempts to comfort the child, while the child is once again free to play; then,

		mother and the stranger leave the room
6	Baby Alone (3 mins or less)	Baby stay alone, then stranger re-enter the room
7	Stranger, Baby (3 mins or less)	the stranger attempts to engage the baby
8	Mother, Baby (3 mins)	Mother re-enter the room, greeting baby, then pick him up. Meanwhile stranger leaves unobtrusively

Figure 2.1 Episodes of Strange Situation Protocol(wiki pedia)

Four aspects of the child's behaviour are observed and analysed:

1. The amount of exploration the child engaged in throughout;
2. The child's reaction to the departure of its parent;
3. The stranger anxiety of child when the child is alone with the stranger
4. The child's reunion behaviour with its caregiver.

2.1.2 Attachment Classification

On the basis of experimental result, Ainsworth grouped and identified three "organised" attachment styles, or patterns: secure, avoidant (insecure), and ambivalent or resistant (insecure). Disorganised attachment was added after further research by Mary Main [10].

The child and caregiver behaviours are categorised four different attachments relationship with the parent in figure 2.2:

Attachment Pattern	Child	Primary Caregiver
Secure	Uses caregiver as a secure base for exploration. Protests caregiver's departure and seeks proximity and is comforted on return, returning to exploration. May be comforted by the stranger but show clear preference for the caregiver. Secure attachment is seen as the most adaptive attachment style.	Responds appropriately, Promptly, consistently to needs. Caregiver has successfully formed a secure parental attachment bond to the child.

Ambivalent/ Resistant	Unable to use caregiver as a secure base, seeking proximity before separation occurs. Distressed on separation with ambivalence, anger, reluctance to warm to caregiver and return to play on return. Concentrated with caregiver's availability, seeking contact but resisting angrily when it is achieved. Not easily calmed by stranger. In this relationship, the child always feels anxious as the caregiver's availability is never consistent.	Inconsistent between appropriate and neglectful responses. Generally will only respond after increased attachment behavior from the infant.
Avoidant	Little affective sharing in play. Little or no distress on departure, little or no visible response to return, ignoring or turning away with no effort to maintain contact if picked up. Treats the stranger similarity to the caregiver. The child feels that there is no attachment; therefore, the child is rebellious and has a lower self-image and self-esteem.	Little or no response to distressed child. Discourages crying and encourages independence.
Disorganised	Cry during separation, but avoid the mother when she returns or may or may approach the mother then freeze or fall to the floor. Some show stereotyped behaviour, rocking. (Lack of coherent attachment strategy shown by contradictory, disoriented behaviours such as approaching but with the b ack turned.)	Frightened or frightening behaviour, intrusiveness, withdrawal, negativity, role confusion, affective communication errors and maltreatment. Very often associated with many for victims of abuse towards the child.

Figure 2.2 Child and caregiver behaviour patterns before the age of 18 months

2.1.3 Implications of Attachment

Ainsworth and her group found the correlation between Strange Situation and behavior at home, the extent of parent responding child at home indicates the sensitively of respond Strange Situation behavior [6]. The further psychometric validity of the methods is found based on Ainsworth work.

The other correlation is focus on differences in attachment to later development of individuals, whose attachment with the primary caregiver was assessed in the Strange Situation at 12 and 18 months of age, shows that the individuals of secure attachment have be found to enjoy a favorable development, being less dependent on adults, having more powers of concentration, and more successful in their peer interactions. [12]

The Strange Situation procedure as an assessment method for individuals difference is proved by the strong correlates established among Strange Situation classifications and past(interactions at home) and future (attachment classification in older age, interaction with peers, state of mind with respect to attachment in adulthood, etc) events in the individuals' life.

2.1.4 Significance of Attachment Theory

The attachment established in infancy that not change afterwards, has a profound effect on their own children's attachment [13][14]. With few exceptions, victims of abuse will result in fear, anxiety, loneliness, emotional lack of support, and being ignored, degraded and humiliated, feeling unloved and unwanted, and being powerless when terrorised or tormented by parents or carers. Iwaniec found that 70% of the parents showed emotional unavailability to their children and lack of communication with them. Indeed, 60% of these mothers reported having had a very poor relationship with their own parents (specially their own mothers) their whole lives, feeling neglected or unwanted [15].

Such high correlation between the attachment styles of parents and their child in parenting interactions, attachment theory provide a mechanism to find how optimal and stable child development patterns can be achieved, and whether adverse attachment styles can be broken between generations.

2.2 Game Theory

2.2.1 Strategic Games (matrix game)

A strategic game is a model of interactive decision-making in which each decision-maker chooses his plan of action once and for all, and these choices are made simultaneously.

A *strategic game* [16] is a tuple $(n, A_{1\ldots n}, R_{1\ldots n})$, where n is the number of players, A_i is the set of actions available to player i (and A is the joint action space $A_1 \times \ldots \times A_n$), and R_i is player i's payoff function $A \rightarrow \boldsymbol{R}$. The players select actions from their available set with the goal to maximize their

payoff, which depends on all the players' actions. These are often called matrix games, since the R_i functions can be written as n -dimensional matrices.

2.2.2 Type of Strategic Games

Two common classes of strategy games are purely collaborative and purely competitive games, which classified according to the structure of their payoff functions.

In purely collaborative games, all agents share same payoff functions, thus one chooses the action in best interest of himself also the best interest of all the agents.

In purely competitive games, which also called zero-sum games, since the two players, where one 's payoff function is the negative of the other (i.e. $R_1 = -R_2$). For example, a common matrix games are in Figure 2.3 shows typical zero-sum games. In these games there are two players; Player 1 is row player, and Player 2 selects a column of the matrix. If the choices are the same then Player 1 takes a dollar from Player 2, otherwise Player 1 gives a dollar to Player 2. The entry of the matrix they jointly select determines the payoffs.

$$\begin{bmatrix} (1,-1) & (-1,1) \\ (-1,1) & (1,-1) \end{bmatrix}$$

$$R_1 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, R_2 = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}$$

Figure 2.3 Matching Pennies Game

Other games, including purely collaborative games, are called general-sum games. The most classic general-sum game is Prison Dilemma. The description of the game and a general form payoff matrix for the prisoner's dilemma game shows in Figure 2.4 and Figure 2.5.
To be a prisoner's dilemma, the following must be true: $B > A > D > C$

	Prisoner B stay silent (cooperates)	Prisoner B betrays (defects)
Prisoner A stay silent (cooperates)	Each serves 1 month	Prisoner A : 1 year Prisoner B: goes free
Prisoner A betrays (defects)	Prisoner A : goes free Prisoner B : 1 year	Each serves 3 months

Figure 2.4 : Description of Prison Dilemma

		Player 2	
		Cooperate	Defect
Player 1	Cooperate	A, A	C, B*
	Defect	B*, C	D*, D*

Figure 2.5 general form of Payoff matrix for the prisoner dilemma game

An interesting property of general sum games is that if there are multiple Nash Equilibria, each of these may have different payoffs. In this case it might be that both players prefer the same equilibrium, or that each player prefers a different equilibrium.

2.2.3 Nash Equilibrium and Pareto Optimality

In game theory, Nash equilibrium (named after John Forbes Nash, who proposed it) is a solution concept of a game involving two or more players, in which each player is assumed to know the equilibrium strategies of the other players, and no player has anything to gain by changing only his own strategy unilaterally. This notion captures a steady state of the play of a strategic game in which each player holds the correct expectation about the other players' behavior and acts rationally. It does not attempt to examine the process by which a steady state is reached.

Definition: A Nash equilibrium of a strategic game $\langle N, (A_i), (\succsim_i) \rangle$ is a profile $a^* \in A$ of actions with the property that for every player $i \in N$ we have
$(a^*_{-i}, a^*_i) \succsim_i (a^*_{-i}, a_i)$ for all $a_i \in A_i$.

Thus for a^* to be a Nash equilibrium it must be that no player r i has an action yielding an outcome that he prefers to that generated when he chooses a^*_i, given that every other player j chooses his equilibrium action a^*_j, Briefly, no player can profitably deviate, given the actions of the other players. Figure 2.6 gives us a concrete example of traditional prisoner dilemma, the (Defect,Defect) is the Nash Equilibrium of the game.

		Player 2	
		Cooperate	Defect
Player 1	Cooperate	3,3	0, 4*
	Defect	4*, 0	1*, 1*

Figure 2.6 An example of payoff matrix of prisoner dilemma game

The concept of Pareto optimality occurs in a number of areas of economics. The allocation of resources in an economy is Pareto optimal, if it is not possible to change the allocation of resources in such a way as to make some people better off without making others worse off. In game theory, a Pareto optimal outcome is one in which no player could be better off without another becoming worse off. A Nash equilibrium, and other outcomes that can be predicted, maynot be Pareto optimal. We say that outcome $a \in S$ is Pareto superior to outcome $b \in S$ if the following both hold:

$$\exists i : u_i(a) > u_i(b)$$

$$\forall i : u_i(a) \geq u_i(b)$$

The Pareto optimal will use for the multiple Nash equilibria of general-sum games.

2.3 Game theoretic Model of The Strange Situation

The Canadian Industrial Problem Solving Workshops Study Group (2006) devised a series of stage game models for the strange situation and use equilibrium to analyse an analysed equilibrium strategies for the mother and child, whereby the mother and child have no incentive to change their strategies. The game theory model was used to attempt to determine how the behavioural characteristics of ‘secure’, ‘ambivalent’ and ‘avoidant’ might emerge for the child as equilibrium choices with respect to the way in which the parent responds to the child's needs. Models focusing on the part of the strange situation at which the parent returns to the room were devised for both one and two-person games. These models form the basis for this entire project and will be referred to throughout, and David concludes an extensive review of the original paper and its conclusions.[1]

2.3.1 The One Person Game

The one person game is based on the action the child chooses (their strategy) under the assumption that the parent's behaviour is independent of the child's. When the parent re- enters the room having left the child alone with the stranger for two minutes, she is modelled as having two possible action choices: they can either attend to the child (`Attend') or completely ignore them (`Ignore').

The mother picks between these actions with probabilities q and $1-q$ respectively, which are normalised to sum to 1.

The scenario is an optimisation problem for the child: they need to pick an action based on the knowledge of the probabilities assigned to the mother's choice of action, such that their anxiety is reduced by the maximum possible amount. It is assumed that the child knows the mother's probabilities based on previous experience of similar stressful situations, and that the child can choose their action from 2 choices: to go to the mother to seek comfort (`Go') or to avoid the mother (`Don't Go'). The payoff matrix for these 2 strategies is shown in Figure 2.7.

		Parent	
		Attend q	Ignore $1-q$
Child	Go	1	$-s$
	Don't Go	0	0

Figure 2.7 payoff matrix for one person(child) game with 2 strategy

2.3.2 The Two Person Game

The two person game model is more practical and rational. In the one person game it assumed that the mother's strategies were fixed and known to the child, however, in real life, there is an game with incomplete information for both players. They learn from every stage of game and reinforce following their own's learning rule to update their information about the game when they do iterated games. Look back the game theoretic model, this two person game forms the basis for games used within this project. In this two person game the mother picks her strategy based on what she thinks the child will do and this is represented as a two-person nonzero-sum, non-cooperative game with the payoff matrix in figure 2.8. The payoff for the mother is depend of the child, which assumes that the gain/loss in stress for the child is also a gain/ loss for her. For example, attending to the child involves some cost c to the mother, hence the payoff for attending when the child picks 'Go' is $1-c$, and the payoff for attending when the child picks `Don't Go' is $-c$.

		Parent	
		Attend (q)	Ignore $(1-q)$
Child	Go (p)	$1\ ,\ 1-c$	$-s,-s$
	Don't-Go $(1-p)$	$0,\ -c$	$0,0$

Figure 2.8: payoff matrix for two person game with 2 strategy

The Nash equlibria for relative (ordinal) payoffs were determined for various combinations of the parameters *s* and *c*. The payoff matrices and resulting equilibrium combination of strategies were grouped into 4 groups according to the inequality conditions placed on the parameters, and are summarized in figure 2.9:

Group1 Games $0 < -s < 1$ & $1-c < -s$	*Type IA*1 : $1-c < 0 < -s < 1$	$\begin{bmatrix}(4,2) & (\mathbf{3}^*,\mathbf{4}^*)\\(2,1) & (2,3)\end{bmatrix}$	NE: *secure* (Go, Ignore)
	*Type IA*2 : $0 < 1-c < -s < 1$	$\begin{bmatrix}(4,3) & (\mathbf{3}^*,\mathbf{4}^*)\\(2,1) & (2,2)\end{bmatrix}$	
Group2 Games $0 < -s < 1$ & $1-c > -s$	*Type IB*:	$\begin{bmatrix}(\mathbf{4}^*,\mathbf{4}^*) & (3,3)\\(2,1) & (2,2)\end{bmatrix}$	NE: *secure* (Go, Attend)
Group3 Games $0 > -s$ & $1-c < -s$	*Type IIA*:	$\begin{bmatrix}(4,2) & (2,3)\\(3,1) & (\mathbf{3}^*,\mathbf{4}^*)\end{bmatrix}$	NE: *avoidant* (Don't Go, Ignore)
Group4 Games $0 > -s$ & $1 - c > -s$	*Type IIB*1*a*: $0 > 1 - c > -s > -c$	$\begin{bmatrix}(\mathbf{4}^*,\mathbf{3}^*) & (2,2)\\(3,1) & (\mathbf{3}^*,\mathbf{4}^*)\end{bmatrix}$	NE: *secure &avoidant* (Go, Attend) & (Don't Go, Ignore)
	*Type IIB*1*b* : $0 > 1 - c > -c > -s$	$\begin{bmatrix}(\mathbf{4}^*,\mathbf{3}^*) & (2,1)\\(3,2) & (\mathbf{3}^*,\mathbf{4}^*)\end{bmatrix}$	
	*Type IIB*2*a* : $1 - c > 0 > -s > -c$	$\begin{bmatrix}(\mathbf{4}^*,\mathbf{4}^*) & (2,2)\\(3,1) & (\mathbf{3}^*,\mathbf{3}^*)\end{bmatrix}$	
	*Type IIB*2*b* : $1 - c > 0 > -c > -s$	$\begin{bmatrix}(\mathbf{4}^*,\mathbf{4}^*) & (2,1)\\(3,2) & (\mathbf{3}^*,\mathbf{3}^*)\end{bmatrix}$	
In addition, a third Nash Equilibrium in *mixed strategies* was identified for games in group 4. These were $p = \frac{c}{1+s}$ for the child, and $q = \frac{s}{1+s}$ for the parent(for c>0 or s>0).			

Figure 2.9: Four Groups Games

In the case of group 4 games, it was argued that the equilibrium in mixed strategies has the qualities of an ambivalent relationship. This is because the probability that the child will seek comfort (p) is proportional to the cost to the mother of attending (c), therefore the more it costs the mother to attend to the child, the more the child will seek comfort. The probability that the mother chooses to attend to the child (q) is proportional to the stress the child suffers as a result of not being comforted (s), therefore the higher this stress the more likely the mother is to attend to the child.

2.4 Models of Disorganised Attachment

To capture disorganised behaviours, the further study on the disorganised attachment by David[1] expanded the original 2x2 game into a 3x3 game such that the parent and child both have 3 action options. In this section it represents two alternative models which each capture a different element of disorganised attachment. These models in attachment strategy exhibits itself in the strange situation through sequential or simultaneous combinations of complex, contradictory behaviours (e.g. strong proximity seeking followed by avoidance or distress), confused expressions, rapid change in affect and/or complete dissociation [10].

2.4.1 Model 1: The `Hostile' Mother

In figure 2.10, a third action `Freeze' is introduced to capture the dissociation observed in disorganised children during the strange situation. Myers has attempt to explain that such dissociation in children provides some temporary relief from extreme trauma (such as that which would result from a negative-intrusive interaction), however in the long term results in a significant decrease in psychological functioning [22]. If the child chooses `Freeze', as the internalised cost associated with the child's emotional disconnection from their environment, there is an associated emotional cost d. The costs associated with being frightened are g and i, and it is assumed that even if the child chooses `Don't Go', their proximity to the parent is such that there is still a negative effect on the child.

		Parent		
		Attend	Ignore	Frighten
Child	Go	1	-s	-g
	Don’t Go	0	0	-i
	Freeze	-d	-d	-d

Figure 2.10 Child’s Payoff Matrix of Model 1

The third action for the mother is `Frighten' presenting a negative-intrusive behaviour. Their payoff matrix is given in Figure 2.11. If the mother chooses `Attend' and the child chooses `Don't Go', then the relief in stress to the child is 0 (as before), however there is an emotional cost to the mother of u, since the child has chosen not to go to the mother, her emotional needs are not fulfilled and there is a negative cost. Indeed, for any case in which the child chooses either `Don't Go' or `Freeze', and the mother chooses to `Attend' or `Frighten', the parent's emotional needs with regards to this role reversal will be unmet.

When the mother chooses `Frighten', there is an emotional cost f associated with this action. As in the original 2x2 game, Cittern assumed that any increase in stress to the child is also an increase in stress to the mother, as (Go, Frighten), (Don't Go, Frighten) representing the feature. However, action pair (Go, Ignore) shows the other feature, which any stress which the child experiences as a result of dissociation is internalised and thus is not passed onto the mother. Therefore, it is assumed that when the child dissociates (freezes) and refrains from making their stress outwardly apparent, that it appears to the mother as if the child's stress has decrease, hence the `1' in the (Freeze, Attend) and (Freeze, Frighten) payoffs.

		Parent		
		Attend	Ignore	Frighten
	Go	1	-s	-g-f
Child	Don’t Go	0-u	0	-i-f-u
	Freeze	1-u	0	1-f-u

Figure 2.11 Parent’s Payoff Matrix of Model 1

The stage game payoff matrix is given in Figure 3.3, where for the child P_v, P_w and (1-P_V - P_w)are the mixed strategy probability labels associated with `Go', `Don't Go' and `Freeze' respectively, and for the mother$R_x$ , $R_y$ and 1-$R_x$-$R_y$ are the probability labels associated with the `Attend', `Ignore' and `Frighten' actions.

		Parent		
		Attend R_x	Ignore R_y	Frighten 1-R_x-R_y
Child	Go (P_v)	1,1	-s,-s	-g,-g-f
	Don't Go (P_w)	0, 0-u	0, 0	-i , -i-f-u
	Freeze (1-P_V - P_w)	-d, 1-u	-d, 0	-d, 1-f-u

Figure 2.12 Stage Game of 'Hostile' Mother

For all games based on this model, some basic assumptions regarding the payoffs were made:

$\mathbf{1 > 0 > -d > -s > -g;}$
i.e. that the payoff to the child of dissociation, rejection and being frightened are all negative and preferred in that order. This assumption has been made specifically to capture cases where dissociation can be the optimal response to interactions in which the child both experiences fear and predicts it as a likely outcome of the interaction.
$\mathbf{1 > 0 > -d > -i > -g;}$
reflects the assumption that the parent and child are in close enough proximity such that even if the child chooses `Don't Go', if the mother has chosen `Frighten' (e.g. with verbal abuse) then this will still have some kind of negative effect on the child, albeit less so than if the child had committed to seek comfort from the mother.
$\mathbf{1 - f - u > -i - f - u;}$
For the mother, we assume that $1 - f - u > -i - f - u$, which follows from 1 > –i (i.e. that the child still experiences some negative effect from the `Frighten' action, even if they choose `Don't Go').
$\mathbf{0 > 0 - u;}$
$0 > 0 - u$ reflects the element of role reversal in the mother, and the negative emotional cost should the child choose not to approach.

$1 > 1 - u > 0 - u$;
$1 > 1 - u > 0 - u$ enforces that the mother views `Don't Go' as more of a rejection than a 'Freeze' action.
$0 > -g - f$;
$-g - f < 0$ and $-i - f - u < 0$, i.e. the physical act of frightening the child has a net negative emotional effect on the mother, and due to the element of role reversal attention seeking this is more pronounced when the child chooses `Don't Go' rather than `Freeze'.

Cittern[1] classified the 16 constrains to model the 'Hostile' Mother profile, Table 2.1 below shows the detailed binary tree and the payoff matrices and resulting equilibrium combination of strategies were grouped into 2 groups according to the inequality conditions placed on the parameters, and are summarized in figure 2.13:

$0 - u < -s$	A	a	$0 - u > -s$
$0 - u < -g - f$	B	b	$0 - u > -g - f$
$0 - u < -i - f - u$	C	c	$0 - u > -i - f - u$
$0 - u < 1 - f - u$	D	d	$0 - u > 1 - f - u$
$1 - u < -s$	E	e	$1 - u > -s$
$1 - u < 0$	F	f	$1 - u > 0$
$1 - u < -g - f$	G	g	$1 - u > -g - f$
$1 - u < -i - f - u$	H	h	$1 - u > -i - f - u$
$1 - u < 1 - f - u$	I	i	$1 - u > 1 - f - u$
$-s < -g - f$	J	j	$-s > -g - f$
$-s < -i - f - u$	K	k	$-s > -i - f - u$
$-s < 1 - f - u$	L	l	$-s > 1 - f - u$
$0 < 1 - f - u$	M	m	$0 > 1 - f - u$
$-g - f < -i - f - u$	N	n	$-g - f > -i - f - u$
$-g - f < 1 - f - u$	O	o	$-g - f > 1 - f - u$
$-s < -i$	P	p	$-s > -i$

Table 2.1: Inequality branches in the binary tree for the `hostile' model

Group	Example	Pure Nash Equilibria
Group 1 (216 games) 'D' branches: $0 - u < 1 - f - u$ 'I ' branches : $1 - u < 1 - f - u$ 'M' branches: $0 < 1 - f - u$ 'O' branches: $-g - f < 1 - f - u$	$(6^*, 8^*)$ (2,3) (1,4) (5,1) $(5^*, 6^*)$ (3,5) (4,2) (4,6) $(4^*, 7^*)$ *type ABCDEFGHIJKLMNO*P	(Go, Attend): secure (Don't Go, Ignore): avoidant (Freeze, Frighten): disorgansied
Group 2 (576 games)	$(6^*, 8^*)$ (2,3) (1,4) (5,1) $(5^*, 7^*)$ (3,5) (4,2) (4,7) $(4^*, 6^*)$ *type ABCDEFGHIJKLmNOP*	(Go, Attend) : secure (Don't Go, Ignore): avoidant

Figure 2.13 Two groups games of 'hostile' Mother model

2.4.2 Model 2: The Affective Communication Error

As detailed earlier, it is a class of behaviours that connected with disorganised attachment closely. In this model, the child has the `Go' and `Don't Go' actions that they had in the original 2x2 game, and in addition they have the `Half Go' action, which has been associated with ambivalence (an action representing a guarded or cautious request for comfort) [1].

If the child chooses to `Half-Go', and the mother chooses `Attend', then the reduction in stress for the child h is less than the reduction in stress which would have been achieved had the child fully committed to the attachment encounter.

		Parent		
		Attend	Half-Attend	Ignore
Child	Go	1	-m	-s
	Half Go	h	0	-t
	Don't Go	0	0	0

Figure 2.13 Child's Payoff Matrix of Model 2

Similarly, the mother can choose to Attend or Ignore as in the 2x2 game, and in addition they can choose to `Half-Attend' which in this context represents some kind of affective communication error. For example, the mother may attend to the child in terms of proximity, but not offer any comfort to the child, thereby confusing the child and increasing their stress. Thus the payoff to the child of (Go, Half-Attend) is −m (their stress is increased). In the case where the child chooses `Half-Go' and the mother chooses `Half-Attend', we have the interesting situation where the child is requesting comfort in a guarded, non-committal manner, and the mother is guarded or contradictory in her offering of comfort to the child. In such a situation we assume that the net effect on the child's stress level is 0 (i.e. their stress is neither increased nor decreased). The child's payoff matrix is given in Figure 2.14.

		Parent		
		Attend	Half-Attend	Ignore
Child	Go	1-c	-m-a	-s
	Half Go	h-c	0-a	-t
	Don't Go	0-c	0-a	0

Figure 2.14 Parent's Payoff Matrix of Model 2

		Parent		
		Attend R_d	Half Attend R_e	Frighten 1-R_d-R_e
Child	Go (P_u)	$1,1-c$	$-m,-m-a$	$-s,-s$
	Don't Go (P_v)	$h,h-c$	$0,0-a$	$-t,-t$
	Freeze (1-P_u - P_v)	$0,0-c$	$0,0-a$	0,0

Figure 2.15 Stage Game of 'affective communication error' model

As before, a binary tree was constructed in order to compute all possible ordinal games. The branch structure of the binary tree is given in Table 3.2.

Basic assumptions on the parameters were again made:

$\mathbf{0-c>0-a>0}$
The cost to the mother of playing Attend is always greater than the cost of 'Half-Attend'.

$\mathbf{1 - c > h - c > 0 - c}$
since $c > 0$ this is reflected in the mother's payoff matrix too with the inequality rule $1 - c > h - c > 0 - c$.
$\mathbf{1 > h > 0 > -t > -s}$
Firstly this enforces the rule that the child's stress will increase more as the result of being ignored if they choose `Go' rather than `Half-Go'. Secondly, the rule states that, given that the mother has chosen `Attend', the payoff to the child of `Go' is greater than `Half-Go', and the payoff from `Half-Go' is greater than `Don't Go'
$\mathbf{0 > -m > -s}$
$0 > -m > -s$ which implies that the child is worse off if they choose to seek comfort and the mother completely ignores them, than they are if they choose `Go' and the mother plays `Half-Attend'.

Cittern[1] also classified the 18 constrains to model the 'affective communication error' profile, Table 2.2 below shows the detailed binary tree and the payoff matrices and resulting equilibrium combination of strategies were grouped into 4 groups according to the inequality conditions placed on the parameters, and are summarized in figure 2.14:

$1 - c < -m - a$	A	a	$1 - c > -m - a$
$1 - c < 0 - a$	B	b	$1 - c > 0 - a$
$1 - c < -s$	C	c	$1 - c > -s$
$1 - c < -t$	D	d	$1 - c > -t$
$1 - c < 0$	E	e	$1 - c > 0$
$h - c < -m - a$	F	f	$h - c > -m - a$
$h - c < 0 - a$	G	g	$h - c > 0 - a$
$h - c < -s$	H	h	$h - c > -s$
$h - c < -t$	I	i	$h - c > -t$
$h - c < 0$	J	j	$h - c > 0$
$0 - c < -m - a$	K	k	$0 - c > -m - a$

$0 - c < -s$	L	l	$0 - c > -s$
$0 - c < -t$	M	m	$0 - c > -t$
$0 - a < -s$	N	n	$0 - a > -s$
$0 - a < -t$	O	o	$0 - a > -t$
$-m - a < -s$	P	p	$-m - a > -s$
$-m - a < -t$	Q	q	$-m - a > -t$
$-m < -t$	R	r	$-m > -t$

Table 2.2: Inequality branches in the binary tree for the 'affective communication error' model

Group	**Example**	**Pure Nash Equilibria**
Group 1 (76 games)		(Don't Go, Ignore): avoidant
Group 2 (29 games) `G' branch: $\mathbf{h - c < -s}$ `K' branch: $\mathbf{0 - c < -m - a}$ `o' branch: $\mathbf{0 - a > -t}$ `p' branch: $\mathbf{-m - a > -s}$	$(\mathbf{6,3})$ $(\mathbf{2,5})$ $(\mathbf{1,4})$ $(\mathbf{5,2})$ $(\mathbf{4^*,7^*})$ $(\mathbf{3,6})$ $(\mathbf{4,1})$ $(\mathbf{4,7})$ $(\mathbf{4^*,8^*})$ $\boldsymbol{type\ ABCDEFGHIJKLMnopQr}$	(Half-Go, Half-Attend): Ambivalent/Disorganised (Don't Go, Ignore) : avoidant
Group3 (169 games) `a' branch: $\mathbf{1 - c > -m - a}$ `c' branch: $\mathbf{1 - c > -s}$ ` g ' branch: $\mathbf{h - c} > 0 - \mathbf{a}$	$(\mathbf{6^*,8^*})$ $(\mathbf{2,3})$ $(\mathbf{1,1})$ $(\mathbf{5,7})$ $(\mathbf{4,5})$ $(\mathbf{3,2})$ $(\mathbf{4,4})$ $(\mathbf{4,5})$ $(\mathbf{4^*,6^*})$ $\boldsymbol{type\ abcdefghiJklmnopq}$	(Go, Attend) : secure (Don't Go, Ignore): avoidant
Group 4 (81 games) `a' branch: $\mathbf{1 - c > -m - a}$ `c' branch: $\mathbf{1 - c > -s}$ ` G ' branch: $\mathbf{h - c < 0 - a}$ `H' branch: $\mathbf{h - c} < -s$ `J' branch: $\mathbf{h - c < 0}$ `n' branch: $\mathbf{0 - a > -s}$ `o' branch: $\mathbf{0 - a > -t}$	$(\mathbf{6^*,8^*})$ $(\mathbf{2,3})$ $(\mathbf{1,4})$ $(\mathbf{5,2})$ $(\mathbf{4^*,6^*})$ $(\mathbf{3,5})$ $(\mathbf{4,1})$ $(\mathbf{4,6})$ $(\mathbf{4^*,7^*})$ $\boldsymbol{type\ abcdeFGHIJKLMnoPQr}$ $(\mathbf{6^*,5^*})$ $(\mathbf{2,3})$ $(\mathbf{1,4})$ $(\mathbf{5,1})$ $(\mathbf{4^*,7^*})$ $(\mathbf{3,6})$ $(\mathbf{4,1})$ $(\mathbf{4,7})$ $(\mathbf{4^*,8^*})$ $\boldsymbol{type\ aBcDEFGHIJKLMnoPQr}$	(Go, Attend) : secure (Half-Go, Half-Attend): Ambivalent/Disorganised (Don't Go, Ignore) : avoidant

Figure 2.14 Four groups games of 'affective communication error' model

CHAPTER 3

Multi-agent Reinforcement Learning

Stochastic games are multi-stage games in which agents' payoff functions may change from stage to stage. Each agent in such a game faces a Markov decision process (MDP), which is interwined with other agent's MDPs. The framework of stochastic games can be used to model a wide range of dynamic multi-agent systems. In extending Q-learning to multi-agent environments, we adopt the framework of general-sum stochastic games. In a stochastic game, each agent's reward depends on the joint action of all agents and the current state, and state transitions obey the Markov property. The stochastic game model includes Markov decision processes as a special case where there is only one agent in the system. General-sum games allow the agents' rewards to be arbitrarily related.

3.1 Stochastic Game

Markov decision processes (MDPs) provide a mathematical framework for modelling decision-making in situations where outcomes are partly random and partly under the control of a decision maker and MDPs are useful for studying a wide range of optimization problems solved via dynamic programming and reinforcement learning. SGs are very natural extension of MDPs to multiple agents. They are also an extension of matrix games to multiple states. Each state in a SG can be viewed as a matrix game with the payoffs for each joint action determined by $R_i\ (s, a)$. After playing discrete matrix game and obtains the correspond rewards the agents are transitioned to another state leading to another matrix games which determined by their joint action. The crucial point is that each agent has its own separate reward function. The dynamic changes among agents gives us a new problem to learn a stationary though to find a possibly stochastic policy $\rho: S \rightarrow PD\ (A_i)$ (i.e. maps states to a probability distribution over its actions). The goal is maximizes the agent's discounted future reward with discounted factor γ.

Definition: A stochastic game is a tuple $(n, S, A_{1\ldots n}, T, R_{1\ldots n})$, where n is the number of players, S is a set of states, A_i is the set of actions available to player i (and A is the joint action space $A_1 \times \ldots \times A_n$), T is a transition function $S \times A \times S \rightarrow [0,1]$, and R_i is player i's reward(payoff) function $S \times A \rightarrow \boldsymbol{R}$.
Stochastic games, which generalize the classical matrix games between players, have been increasingly employed in multi-agent systems to model learning behaviour.

Given state s, agents independently choose actions $a_1 \dots a_n$ (n-agents' actions). and receive rewards $R_i(s, a_1 \dots a_n)$,

i =1… n. The state then transits to the next state s' based on fixed transition probabilities, satisfying the constraint

$$\sum_{s' \in S} p(s'|s, a_1 \dots a_n) = 1.$$

3.1.1 Example of Stochastic Games

We introduces an example of stochastic game described by Hu[24]. She deals with grid-world games, which are also two-player dynamic games described in figure 3.1. In both games, two agents start from respective lower corners, trying to reach their goal cells in the top row. An agent can move only one cell a time, and the action space of agent i, i = 1,2, is A_i ={Left, Right, Down, Up}. The state space is S = {(0 1),(0 2),…,(8 7)}, where a state s = (l_1, l_2) represents the agents' joint location. Agent i's location l_i is represented by position index, as shown in Figure 3.1 as well**.** If two agents attempt to move into the same cell (excluding a goal cell), they are bounced back to their previous cells, with a negative payoff -1. The game continues until an agent reaches its goal. Reaching the goal earns a positive reward, set 100. Otherwise, gain no payoff. In case both agents reach their goal cells at the same time, both are rewarded with positive payoffs 100.

Let $L(l, a)$ be the potential new location resulting from choosing action a in position l.The reward function is, for i = 1,2

$$r_i = \begin{cases} 100 & if\ L(l_i, a_i) = Goal_i \\ -1 & if\ L(l_1, a_1) = L(l_2, a_2)\ and\ L(l_2, a_2) \neq\ Goal_j.\ j = 1,2 \\ 0 & otherwise \end{cases}$$

The state transitions are deterministic in left one (Grid Game 1). In right one (Grid Game 2), state transitions are deterministic except the following: if an agent chooses *Up* from position 0 or 2, it moves up with probability 0.5 and remains in its previous position with probability 0.5. Thus, when both agents choose action *Up* from state (0 2), the next state is equally likely (probability 0.25 each) to be (0 2), (3 2), (0 5), or (3 5). When agent 1 chooses *Up* and agent 2 chooses *Left* from state (0 2), the probabilities for reaching the new states are: P((0 1)|(0 2), *Up*, *Left*) = 0.5, and P((3 1)|(0 2), *Up*, *Left*) = 0.5. Similarly, we have P((1 2)|(0 2), *Right*, *Up*) = 0.5, and P((1 5)|(0 2), *Right*, *Up*) = 0.5.

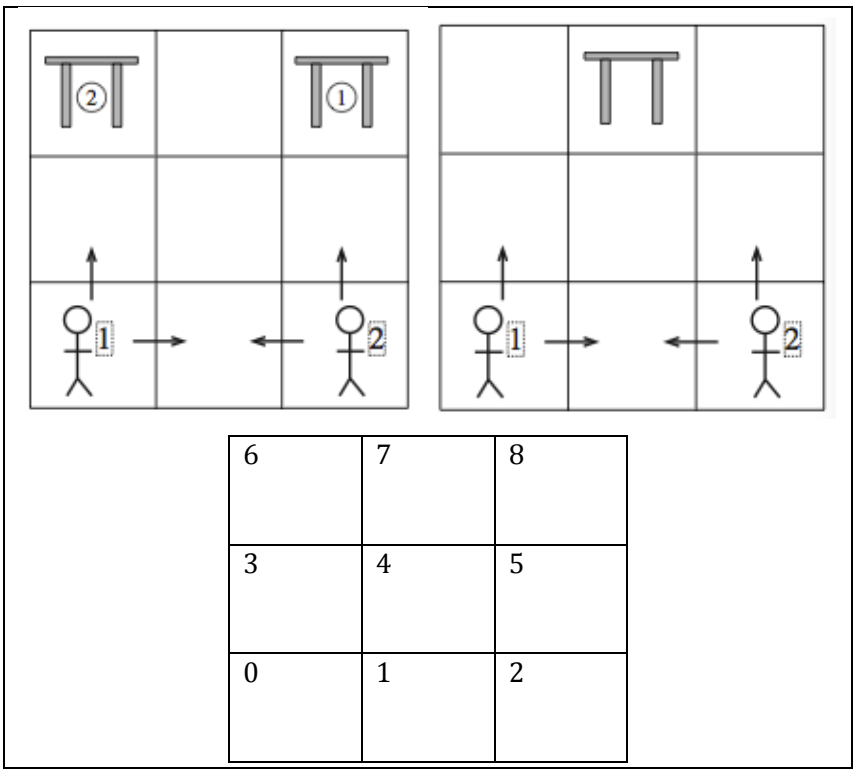
The learning algorithms of Nash Q-value have two crucial variables: value of the game and Nash Q – value.

The value of the game for agent 1 for example, is defined as its accumulated reward when both agents follow their Nash equilibrium strategies,(π_1^*, π_2^*)

$$v_1(s_0) = \sum_t \beta^t E(r_t | \pi_1^*, \pi_2^*, s_0).$$

Based on the values for each state, we can derive the Nash Q-values for agent 1 in state s_0,

$$Q_1(s_0, a_1, a_2) = r_1(s_0, a_1, a_2) + \beta \sum_{s'} p(s'|s_0, a_1, a_2)\, v_1(s')$$



6	7	8
3	4	5
0	1	2

Figure 3.1 Two grid-world Games and location index for grid games

3.1.2 Parent-Child Game in Stochastic Game Framework

An example of stochastic games setting and probability of state transition are described above. For a simple 2X2 games, $\begin{bmatrix}(4,2) & (2,3)\\ (3,1) & (\mathbf{3}^*,\mathbf{4}^*)\end{bmatrix}$ parent-child interactions gam. For child A_1 ={Go, Don't Go}, for parent, A_2 ={Attend, Ignore}. The state space for child in this game is S_1 = {(4,2,3,3),(4,3,2,2)…}, which contains 9(= 3!). And the state space of parent S_2 ={(2,3,1,4), (1,2,3,4)….}, which contains 24 (= 4!) different states. The reward in original game is the payoff value in the matrix stage game. For example, if the joint actions are (Go, Attend), the reward for parent is 2, for child it is 4.

Using fixed transition probabilities in grid game make senses to the game setting. In our project, the fixed transition probabilities cannot set in advance; the probabilities are learning also in a dynamic way, the section 3.4 explains the action chosen rules we use as the state transition function.

$$P(a_i|s) = \frac{k^{Q(s,a_i)}}{\sum_i k^{Q(s,a_i)}}$$

However, the only drawbacks are that states of both agents are not considered simultaneously and when calculating the probability of action for next round also not considering at joint action level. The states of both agents are easy to solve, but the joint actions is need a complex nash equilibrium method to analyse. In this project, an intuitive method, reinforcement rule is used replaced the nash equilibrium method as refinement guiding realization of game goal. So the stochastic game framework is not that standard as Hu's.

3.2 Q-learning in Multiagent Game

The simple Q-learning Cittern used is also used in this project, and the only change we made was let both side of dyads use Q-learning simultaneously, and each side has the same learning ability to make choice to adapt change of others, which we call multiagent Q-learning. The algorithm of Q learning originally proposed by Watkins [2].

The algorithm aims to learn the optimal action choice for each state of the MDP (the policy) for each agent. One of its main strengths is the ability to solve the MDP without the need to know or compute the state transition probabilities. Firstly we define the Q function $Q: S \times A \rightarrow \boldsymbol{R}$. This function calculates a Q value for each action associated with a particular state. Q values are an estimate of the expected reward that a learner (agent) will receive from choosing this action, and so the Q values for a particular state can be thought of as representing the 'quality' associated with each action available to the player at any point in time. These Q values are initialised according to a predefined initialisation rule, and then subsequently updated each time a decision is made and a transition from state s to some other state s' occurs within the MDP. The update rule for the Q values is therefore the learning process by which the player re-evaluates the quality of each state-action combination, according to the rewards received from action choices they make. For example, given that the player has just chosen action a in state s and observed the transition into the new state s' (i.e. $s \xrightarrow{a} s'$), the update rule for the Q value for action a associated with state s is given by:

$$Q(s,a) \leftarrow Q(s,a) + l[R_a(s,s') + \delta \max_{a'} Q(s',a') - Q(s,a)]$$

The parameters for the learning algorithm are the discount factor δ and the learning rate l, where $0 \leq \delta < 1$ and $0 < l \leq 1$.

The discount factor signifies the relative values that the player places on immediate/future rewards - a discount factor of 0 would apply where only current rewards are valued, whilst a discount factor approaching 1 signifies a value placed on long-term reward.

The learning rate determines the extend to which newly acquired information overrides old information- a learning rate approaching 0 implies the decision maker not learning anything, whilst a learning rate of 1 would lead to the learner only considering the most recent information in their state-action quality re-evaluation.

The learning rate can be a constant, however Watkins and Dayan showed that if l takes decreasing successive values l_1, l_2, l_3 .... such that $\sum_{i=1}^{\infty} l_i = \infty$ and $\sum_{i=1}^{\infty} l_i^2 < \infty$, then Q values will converge on the optimal policy if each state-action pair (s,a) is chosen an infinite number of times [7]. In our implementation the rewards associated with each pair (s,a) are based on reinforcements of the agent's payoff matrix and so they are unbounded, however we will follow the typical convention by

associating a learning rate with each state action pair and set it according to $l(s,a) = \frac{1}{n(s,a)}$, where $n(s,a)$ is the number of times action a in state s.

3.3 Properties of Multi-agent Learning Algorithms

Michael Bowling and Manuela Veloso contributes two crucial properties of learning algorithms for stochastic games:

Property 1 *(Rationality) If the other players' policies converge to stationary policies then the learning algorithm will converge to a policy that is a best-response to their policies.*

This is a fairly basic property requiring the player to behave optimally when the other players play stationary strategies. This requires the player to learn a best-response policy in this case where one indeed exists. Algorithms that are not rational often opt to learn some policy independent of the other players' policies, such as their part of some equilibrium solution. This completely fails in games with multiple equilibria where the agents cannot *independently select* and play equilibrium.

Property 2 *(Convergence) The learner will necessarily converge to a stationary policy. This property will usually be conditioned on the other agents using an algorithm from some class of learning algorithms.*

The second property requires that, against some class of other players' learning algorithms (ideally a class encompassing most "useful" algorithms), the learner's policy will converge. For example, one might refer to convergence with respect to players with stationary policies, or convergence with respect to rational players.

In this paper, we focus on convergence in the case of self- play. That is, if all the players use the same learning algorithm do the players' policies converge? This is a crucial and difficult step towards convergence against more general classes of players. In addition, ignoring the possibility of self-play makes the naive assumption that other players are inferior since they cannot be using an identical algorithm.

In combination, these two properties guarantee that the learner will converge to a stationary strategy that is optimal given the play of the other players. There is also a connection between these properties and Nash equilibria. When all players are rational, if they converge, then they must have converged to Nash equilibrium. Since all players converge to a stationary policy, each player, being rational, must con- verge to a best response to their policies. Since this is true of each player, his or her policies by definition must be equilibrium. In addition, if all players are rational and convergent with respect to the other players' algorithms, then convergence to Nash equilibrium is guaranteed.

3.4 Experimentation Strategies

Notice the algorithm of section 3.2 does not specify how the agent chooses actions. One obvious strategy would be for the agent in state s to select the action a that maximizes $Q(s,a)$, thereby exploiting its current approximation Q. However, with this strategy the agent runs the risk that it will overcommit to actions that are found during early training to have high Q values, while failing to explore other actions that have even higher values. In fact, the convergence theorem above requires that each state-action transition occur infinitely often. This will clearly not occur if the agent always selects actions that maximize its current *Q(s,a).* For this reason, it is common in Q learning to use a probabilistic approach to selecting actions. Actions with higher Q values are assigned higher probabilities, but every action is assigned a nonzero probability. One way to assign such probabilities is

$$P(a_i|s) = \frac{k^{Q(s,a_i)}}{\sum_i k^{Q(s,a_i)}}$$

As k increases, the probability of selecting those actions with low Q values becomes smaller, and so one way of looking at the exploration parameter is to say that larger values of k are representative of a player who has more `embedded' behaviour. In this project we will use an exploration parameter of k = 2 and will not look closely at the effect of varying k, which is left as a suggestion for further work.

3.5 other relevant algorithms

This section lists a number of different algorithms for solving stochastic games from both the game theory and reinforcement learning. The algorithms have different assumptions, based on different available model and other agents' specific behavior and control. These algorithms also have their characteristics and advantages; it is an inspiration for the further work to develop these algorithms to model this Strange Situation Protocol or other models.

The Nash- Q algorithms of Hu and Wellman[24] used in general-sum stochastic games shows one crucial differ compared with standard single-agent Q-learning: how to use the Q-values of the next state to update those of the current state. Multi-agent Q-learning algorithm updates with future Nash equilibrium payoffs, whereas single-agent Q-leaning updates are based on the agent's own maximum payoff. In order to learn these Nash equilibrium payoffs, the agent must observe not only its own reward, but those of others as well. The difference between our simple multi-agent algorithms and Hu's algorithms is that we did not

3.5.1 Solution From Game Theory

In game theory, a stochastic game, introduced by Lloyd Shapley in the early 1950s, is a dynamic game with probabilistic transitions played by one or more players. The game is played in a sequence of stages. At the beginning of each stage the game is in some state. The players select actions and each player receives a payoff that depends on the current state and the chosen actions. The game then moves to a new random state whose distribution depends on the previous state and the actions chosen by the players. The procedure is repeated at the new state and play continues for a finite or infinite number of stages. The total payoff to a player is often taken to be the discounted sum of the stage payoffs or the limit inferior of the averages of the stage payoffs.

There are two algorithms that learn a value function over states, $V(s)$. The goal is for V to converge to the optimal value function V^*, which is the expected discounted future reward if the players followed the Nash Equilibrium of the games. $Value(Matrix\ Game)$ and $Solve_i\ (MatrixGame)$ to refer to algorithm for solving matrix game. $Value$ returns the expected value of playing the matrix game's equilibrium and $Solve_i$ returns player i's equilibrium strategy.

Shapley's algorithm shown in table 3.1, is an extension of value iteration to stochastic game. The algorithm uses a temporal differencing technique to backup values of next states into a simple matrix game, $G_s\ (V)$. Update the value function V by solving the matrix game at each state.

1. Initialize V arbitrarily
2. Repeat,
   (a) For each state, $s \in S$, compute the matrix,

$$G_s\ (V) = [g_{a \in A:}\ R(s,a) + \gamma \sum_{s' \in S} T(s,a,s')V(s')]$$

   (b) For each state, $s \in S$, update V,

$$V(s) \leftarrow Value[G_s\ (V)]$$

Table 3.1: Shapley' s algorithm

Pollatschek & Avi-Itzhak is an extension of policy iteration to stochastic game [17]. Table 3.2 shows the algorithm. Each payer selects the equilibrium policy according to the current value function, making use of the same temporal differencing matrix, $G_s\ (V)$, as in shapley's algorithm. The value function is then updated based on the actual rewards of following these policies. This algorithm also computes the equilibrium value function, from which can be derived the equilibrium policies. The convergence of the algorithm can possessed if the transition function T and discount factor γ satisfy some specified situation.

1. Initialize V arbitrarily,

2. Repeat,

$$\rho_i \leftarrow Solve_i\,[G_s\,(V)]$$

$$V(s) \leftarrow E\{\sum \gamma^t r_t | s_0 = s\,, \rho_i\}$$

where s_0 is the initial state, r_t is the reward at time t, and $\gamma \in [0,1)$ is the discount factor

Table 3.2: Pollatschek & Avi-Itzhak algorithm

3.5.2 Solutions From RL

Reinforcement learning solutions take a different approach to finding policies. It is generally assumed the model of the world (T and R) are not known but must be observed through experience. The agents are required to act in the environment in order to gain observations of T and R. The other distinguishing characteristic is that these algorithms focus on the behaviour of a single agent, and seek to find the equilibrium policy for that agent.

Nash-Q

Hu & Wellman [19] extended the Minimax-Q algorithm to general-sum games. The algorithm is designed to directly learn Nash equilibrium. Table 3.3 gives the structure of the algorithms. The extension requires that each agent maintain Q values for all the other agents. Also, the linear programming solution used to find the equilibrium of zero-sum games is replaced with the quadratic programming solution for finding equilibrium in general-sum games.

In their report

This algorithm is the first to address the complex problem of general-sum games. But their algorithm requires a number of very limiting assumptions. The most restrictive of which limits the structure of all the intermediate matrix games faced while learning (i.e. $Q(s,a).$)The largest difficulty is that it is impossible to predict whether this assumption will remain satisfied while learning. [20]

The other assumption to note is that the game must have a unique equilibrium, which is not always true of general-sum stochastic games. This is necessary since the algorithm strives for the opponent-independence property of Minimax-Q, which allows the algorithm to converge almost regardless of the other agent's actions. With multiple equilibria it's important for all the agents to play the same equilibrium in order for it to have its reinforcing properties. So, learning independently is not possible.

1. Initialize $Q(s \in S, a \in A)$ arbitrarily, and set α to be the learning rate.

2. Repeat,

(a) From state s select action a_i that solves the matrix game $[Q(s,a)_{a \in A}]$, with some exploration.

(b) Observing joint-action a, reward r, and next state s',

$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha\,(r + \gamma\, V(s'))$,

where,

$V(s) = Value\,([Q(s,a)_{a \in A}])$.

Table 3.3 Nash-Q algorithm

Policy Hill Climbing

Policy Hill Climbing is another extension of Q- learning to play mixed policies but does not converge in experiment, show below in table 3.4.

1. Let α and δ be learning rates. Initialize,

$Q(s,a) \leftarrow 0, \quad \pi(s,a) \leftarrow \frac{1}{|A_i|}$.

2. Repeat,

(a) From state s select action a with probability $\pi(s,a)$ with some exploration.

(b) Observing reward r and next state s',

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha\,(r + \gamma \max_{a'} Q(s',a')).$$

(c) Update $\pi(s,a)$ and constrain it to a legal probability distribution,

$$\pi(s,a) \leftarrow \pi(s,a) + \begin{cases} \delta; & if\ = argmax_{a'} Q(s,a') \\ \frac{-\delta}{|A_i| - 1}; & otherwise \end{cases}$$

Table 3.4: Policy hill-climbing algorithm (PHC) for player i

WoLF Policy Hill-Climbing

WoLF Policy Hill-Climbing is based on a simple principle: "learning quickly while losing, slowly while winning." It solves the naïve policy hill-climbing algorithm, which cannot converge in experiment. The basic idea is to vary the learning rate used in such way to encourage convergence, without sacrificing rationality. The two learning rates δ_l and δ_w is used to update the policy depends on whether the agent is currently winning (δ_w) or losing (δ_l). This is determined by

comparing the expected value, using current Q-value estimates, of following the current strategy or policy π in the current state with that of following the average policy $\bar{\pi}$. If the expectation of the current policy is smaller (i.e. the agent is "losing") then the larger learning rate, δ_l is used.

1. Let α, $\delta_l > \delta_w$ be learning rates. Initialize,

$$Q(s,a) \leftarrow 0, \quad \pi(s,a) \leftarrow \frac{1}{|A_i|}, \quad C(s) \leftarrow 0$$

2. Repeat,

(a) From state s select action a with probability $\pi(s,a)$ with some exploration.

(b) Observing reward r and next state s',

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha\,(r + \gamma \max_{a'} Q(s',a')).$$

(c) Update estimate of average policy, $\bar{\pi}$,

$$C(s) \leftarrow C(s) + 1$$

$$\forall\, a' \in A_i, \quad \bar{\pi}(s,a') \leftarrow \bar{\pi}(s,a') + \frac{1}{C(s)}(\pi(s,a) - \bar{\pi}(s,a'))$$

(d) Update $\pi(s,a)$ and constrain it to a legal probability distribution,

$$\pi(s,a) \leftarrow \pi(s,a) + \begin{cases} \delta; & if\ = argmax_{a'} Q(s,a') \\ \dfrac{-\delta}{|A_i| - 1}; & otherwise \end{cases}$$

Where

$$\delta = \begin{cases} \delta_w \quad ; & \sum_a \pi(s,a)Q(s,a) > \sum_a \bar{\pi}(s,a)Q(s,a) \\ \delta_l; & otherwise \end{cases}$$

Table 3.5 : WoLF policy hill-climbing algorithm for player i

CHAPTER 4

Iterated Game and Simulation

In extending the single stage game to an iterated game the behaviours and patterns of agents would be apparent gradually. The iterated game can model a change from one attachment style to another. As mentioned in background, classification of the relationship of parents and children would be not stable in long term, such as stressful situations such as divorce, illness or abuse, or a drastic change in the way in which the caregiver interacts with the child. To the contrary, the opposite transition has also observed in cases where the parent's living conditions have improved and influenced a more positive change in behavior. In this chapter, we focus on the positive transition from insecure to secure attachment.

In games where there are multiple Nash equilibria in pure strategies, it may be the case that the attachment style can move from an insecure to secure relationship simply by moving from one equilibrium to another. Cittern defined two game theoretic strategies for the child: `Best Response To Last Move' (BRTLM) and `Maximisation of Expected Payoff' (MEP), and have shown how this evolution in attachment style can occur even if the child is unaware of any changes to the underlying payoff matrix and continues to play a standard iterated game; using the initial payoff matrix and simply picking their actions according to their game theoretic strategy. In this project, we follow Cittern's simple Q-learning algorithms and experiment strategy for both agents of game and also show how this evolution in attachment style can occur when child and parent both can learning and being guided through changes to the underlying payoff matrix.

The work of David Cittern includes game theoretic model and parts of his repeated game programming code has used in our implementation. In this section, we will observe the game model in quantify experimental level but not in reality experimental level.

4.1 Iterated Game Implementation

An iterated game consists of some number of repetitions of the underlying stage game, where each repetition is called a `round', and each player's overall score is the cumulative sum of their payoffs in each of these rounds. In such a scenario the players will have to take into account the impact of their action in the current round on the behaviour of their opponent in the future, which will be made up of either a known, finite number of rounds or an unknown (possibly infinite) number of rounds. In an iterated game scenario, `strategy' is often used to refer to the rule by which the player chooses their action in each round, rather than a single action choice. Each player may or may not

have access to the moves played in previous rounds of the game, which can be used in order to form some prediction of how their opponent will act.

4.2 Reinforcements

The agents both parent and child's payoff matrix must change over the course of the iterated game in order for new Nash equilibria corresponding to secure attachment to emerge. This means that the both sides of game must change the value they place on individual game outcomes, so that outcomes resulting in the parent choosing 'Attend' and the child choosing 'Go' gradually come to be preferred over avoidant and disorganised interaction outcomes.

The goal is to model a learning process for parent and child, whereby the reinforcement of certain desirable action combinations gradually leads the parent-child dyads into a stable pattern of play which is a secure relationship in this section.

We assume that this reinforcement comes from an external source, such as psychotherapist, who monitors the iterated game and the actions of the agents, by encouraging and praising certain outcomes, gradually influences the relative value that the parent places on each interaction outcome. Thus our data analysis is about the average number of rounds for dyads to reach the expected stable states.

4.2.1 Reinforcement Rule

Only desirable or helpful outcomes to each stage game should be praised and encouraged. We label this set of action-combinations that will trigger reinforcements in the parent and child's payoff matrix. Let the multi-agent game from avoidant attachment as an example, In any single stage game the action-combination (Go, Attend) is desirable, since we are hoping to guide the iterated game such that every game ultimately has this outcome. However, this is not necessarily the only action-combination that should be reinforced: we intuitively observe that the action-combination (Don't Go, Attend) is also a helpful outcome, since the parent attending could encourage the child to `Go' in a following round. Conversely, the outcomes (Go, Ignore) and (Don't Go, Ignore) are neither desirable nor helpful outcomes for parent to any stage game since they will encourage and reinforce `Ignore' behaviours in the parent, and are never praised. At the same time, the child follows the same principle, (Go, Attend) is obviously desirable, $\{(\text{Go}, \text{Attend}), (\text{Go}, \text{Ignore})\}$ is a helpful reinforcement part as well.

As positive reinforcement, define r > 1 to be the reinforcement parameter. Reinforcements are multiplicative: each time a desirable action-combination $(a_c, a_p) \in \eta$ is observed following a round in the iterated game, the corresponding payoff element in the parent's payoff matrix is reinforced by this factor r. For example, if we had experienced 10 rounds of play, where the action combination (Go, Attend) had occurred 3 times, (Don't Go, Attend) had occurred 4 times and (Don't Go, Ignore) 3 times, the parent reinforcement rule is η ={(Go, Attend), (Don't Go, Attend)} then the

parent's payoff matrix will have been reinforced to that in Figure 4.1.1 (a different change occurs for child which reinforcement rule η = {(Go, Attend), (Go, Ignore)}, see figure 4.1.2) The payoff element u has not been reinforced since the action-combination (Go, Ignore) has not occurred, and even if it had it would not be reinforced since (Go, Ignore) $\notin \eta$ of parent. Likewise, the element w has not been reinforced since (Don't Go, Ignore) $\notin \eta$ of parent. (The payoff element p has been reinforced since (Go, Ignore) $\epsilon \eta$ of child, while element q has not been reinforced since (Don't Go, Attend) $\notin \eta$ of child). The ordinal equivalence of this reinforced payoff matrix corresponds to the current state of the parent's underlying MDP (following the 10 rounds of play) and it is therefore these payoff reinforcements that directly cause the state transitions in the MDP.

$$\begin{bmatrix} tr^3 & u \\ vr^4 & w \end{bmatrix}$$

Figure 4.1.1: Example of a reinforced payoff matrix for parent in 2X2 game

$$\begin{bmatrix} or^3 & pr^3 \\ q & y \end{bmatrix}$$

Figure 4.1.2: Example of a reinforced payoff matrix for child in 2X2 game

For 3X3 game, 'hostile' mother profile, the reinforcement :
for parent is η = {(Go, Attend), (Don't Go, Attend), (Freeze, Attend)} ,
for child it is η = {(Go, Attend), (Go, Ignore), (Go, Frighten)} } ;
Under these reinforcement rules, for example, if we had experienced 12 rounds of play where the action combination (Go, Attend) had occurred 3 times, (Don't Go, Attend) had occurred 4 times and (Go, Frighten) 3 times, (Freeze, Attend) for 2 times. These reinforced payoff matrix show in figure 4.1.3, and 4.1.4. (The reinforced mechanism is similar as 2X2 multi-agent games)

$$\begin{bmatrix} tr^3 & u & a \\ vr^4 & w & b \\ cr^2 & d & e \end{bmatrix}$$

Figure 4.1.3: Example of a reinforced payoff matrix for parent in 3X3 game

$$\begin{bmatrix} or^3 & p & fr^3 \\ q & y & g \\ k & l & m \end{bmatrix}$$

Figure 4.1.4: Example of a reinforced payoff matrix for child in 3X3 game

4.3 Simulation Results

In this chapter we have taken the first steps in considering how the attachment style can change at an individual parent-child dyad. We have outlined a model of payoff reinforcement and learning for

the parent, whereby certain outcomes to rounds of the iterated game are encouraged (reinforced) whilst others are not.

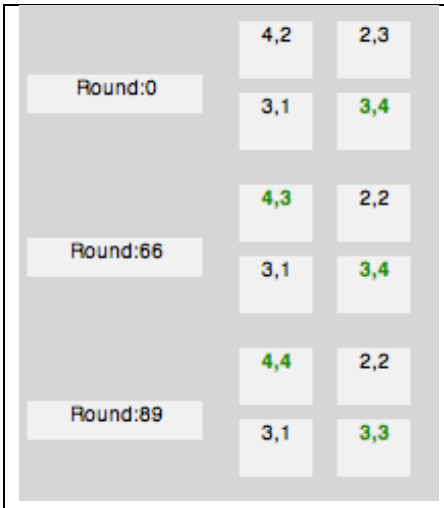
We have defined a number of parameters which can be used to differentiate between types of parents, including the payoff reinforcement rate r (specifying the magnitude of payoff updates), a discount factor δ (to specify the relative values a parent places on immediate and future rewards) and an action-exploration parameter k (which specifies the extent to which non-optimal actions are chosen, or alternatively how embedded the parent is in their current behaviour).

4.3.1 Single Iterated Game

Each individual simulation of iterated game consisted of enough large numbers of rounds that can make sure the convergence can be succeed within the great number of rounds.

For this game, we can set any initial Q values for game to start. So, in every model we follow the same random method for Q values, for example, for a 2X2 game, since the game begins with an avoidant attachment style, the Q values are initialised such that the parent's expectation of rewards associated with the action in each state corresponds to an expectation that the child will play `Don't Go', and without a consideration for any $\begin{bmatrix}2 & 3\\1 & 4\end{bmatrix}$, the initial Q value for the action `Attend' is 1 and for `Ignore' it is 4. Similarly, for the state $\begin{bmatrix}4 & 3\\1 & 2\end{bmatrix}$ the initial Q value for `Attend' is 1 and for `Ignore' it is 2. Next, we assume that in each round of the game the parent and child chooses their action according to the action selection rule mentioned in section 3.4, resulting in an action combination $((a_c, a_p)$. The reward component of the Q value update rule, $R_{a_p}(s, s')$, corresponds to the payoff the parent receives from the action-combination (a_c, a_p). If $(a_c, a_p) \in \eta$ then this reward is a reinforcement of the corresponding payoff in their current payoff matrix, and the parent's ordinal payoff matrix may transition into a new state as a result of this reinforcement. If $(a_c, a_p) \notin \eta$ then the reward for the parent is a non-reinforced payoff and no state transition will occur, i.e. (s, s'). The $R_{a_c}(s, s')$, corresponds to the payoff the child received from the action-combination (a_c, a_p) will following the same setting and rules.

Group3 Games $0 > -s$ & $1 - c < -s$	*Type IIA*:	$\begin{bmatrix}(4,2) & (2,3)\\(3,1) & (\mathbf{3}^*, \mathbf{4}^*)\end{bmatrix}$	NE: *avoidant* (Don't Go, Ignore)



Case 1: 1< r < 3/2 (r = 1.2)

we will consider in detail the state transitions for the case where $\boldsymbol{\eta} = \{(\mathbf{Go}, \mathbf{Attend})\}$. For Type IIA games, the parent's initial payoff matrix (and therefore the starting state for the game) is $\begin{bmatrix} 2 & 3 \\ 1 & 4 \end{bmatrix}$, the final state is $\begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix}$ which corresponds to a Type IIB2a game.

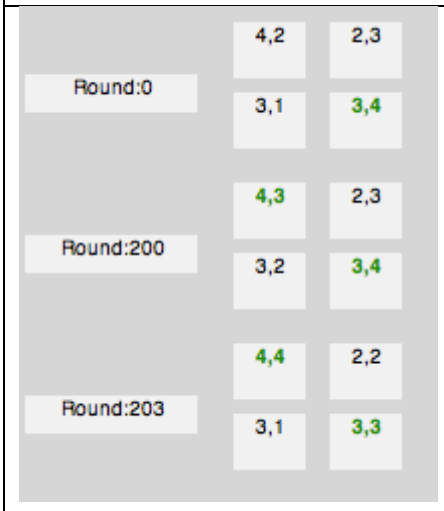
The child's initial payoff matrix is $\begin{bmatrix} 4 & 2 \\ 3 & 3 \end{bmatrix}$, the final state is $\begin{bmatrix} 4 & 2 \\ 3 & 3 \end{bmatrix}$.

Child's state transition for the case where $\boldsymbol{\eta} = \{(\mathbf{Go}, \mathbf{Attend})\}$ has no change. but it will be change

Game state transitions:

$Type\ IIA \rightarrow Type\ IIB1a \rightarrow Type\ IIB2a$

The final evolving matrix NE: secure attachment and avoidant attachment. Secure attachment also the pareto optimality.



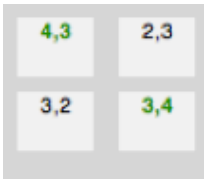
Case 2: r = 3/2

Similar with Case 1

The final evolving matrix NE: secure attachment and avoidant attachment. Secure attachment also the pareto optimality

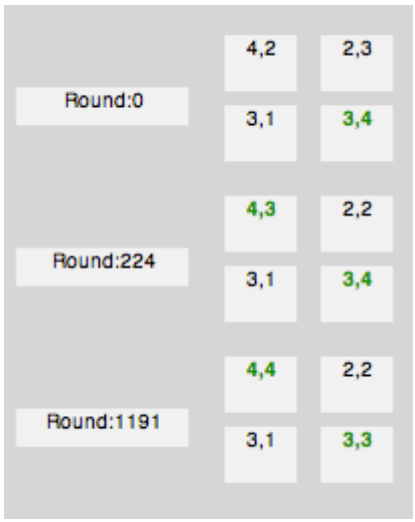
$$\begin{bmatrix} (1,1-c) & (-s,-s) \\ (0,-c) & (0,0) \end{bmatrix}$$

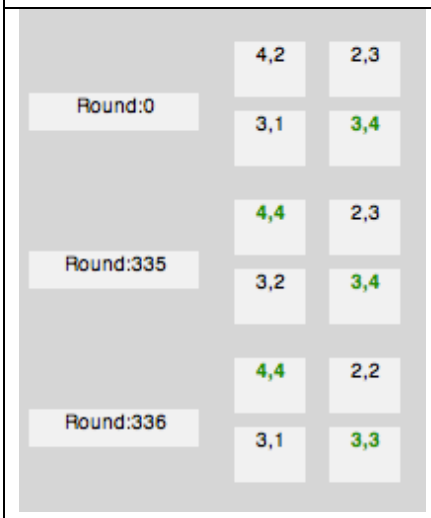
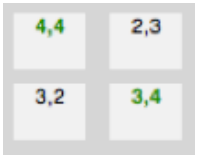
The intermit states :



corresponds $Type\ IIB1a$ which constrains is $0 > 1 - c > -s > -c$

$$Type\ IIA \rightarrow Type\ IIB1a \rightarrow Type\ IIB2a$$

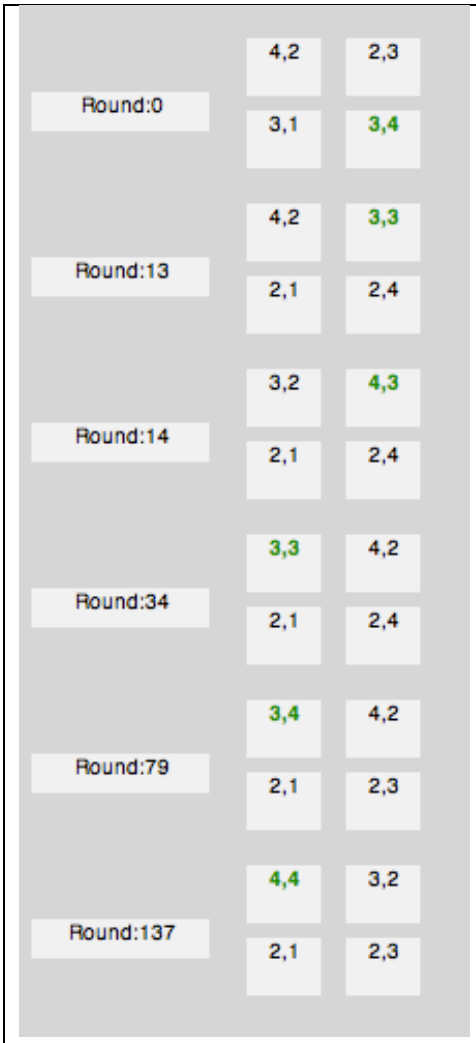
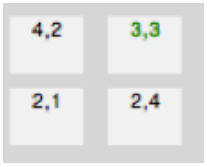
<table>
<tr>
<td> 
</td>
<td>Case 3: $3/2 < r < 2$

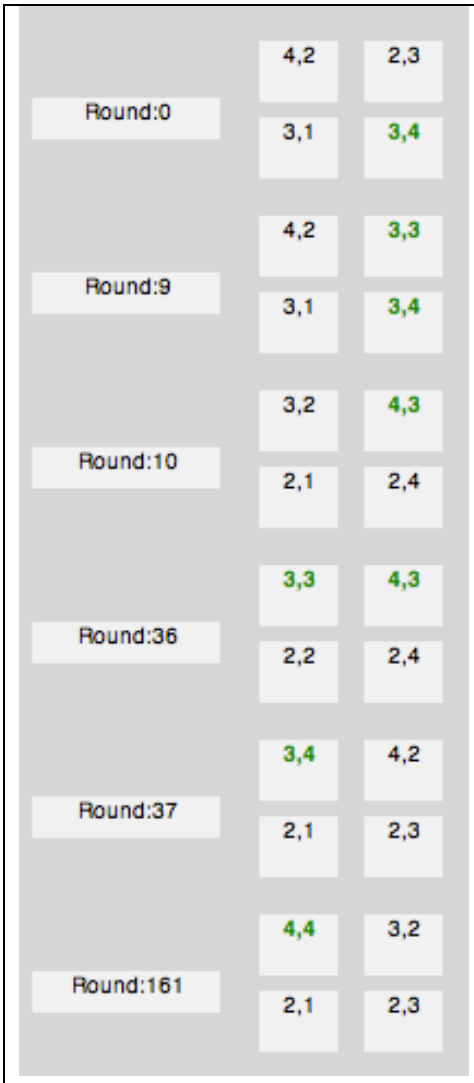
Similar with Case 1, but need more evolving rounds.
Game state transitions:
$Type\ IIA \rightarrow Type\ IIB1a \rightarrow Type\ IIB2a$</td>
</tr>
<tr>
<td> 
</td>
<td>Case 4: $r = 2$
The intermit state in transitions:
 

corresponds to $Type\ IIB2a$ which constrains is $1 - c > 0 > -s > -c$
The NE is (Go, secure), (Don't Go, Ignore).
The state transitions:
$Type\ IIA \rightarrow Type\ IIB2a \rightarrow Type\ IIB2a$
The final evolving matrix NE: secure attachment and avoidant attachment. Secure attachment also the pareto optimality</td>
</tr>
</table>

Figure 4.2 evolving ordinal payoff matrix of avoidant game

Avoidant Attachment for reinforcement rule are $\eta = \{(\mathbf{Go}, \mathbf{Attend})\}$ for agents of game

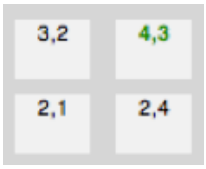
	**Case 5: 1< r < 3/2 (r = 1.2)** we will consider in detail the state transitions for the case where $\boldsymbol{\eta} = \{\{(\text{Go}, \text{Attend}), (\text{Go}, \text{Ignore})\}\}$. For Type IIA games, the parent's initial payoff matrix (and therefore the starting state for the game) is $\begin{bmatrix} 2 & 3 \\ 1 & 4 \end{bmatrix}$, the final state is $\begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix}$. The child's initial payoff matrix is $\begin{bmatrix} 4 & 2 \\ 3 & 3 \end{bmatrix}$, the final state is $\begin{bmatrix} 4 & 3 \\ 2 & 2 \end{bmatrix}$. The intermit state is  correspond to Group 1($Type\ IA2$) which NE is (Go, Ignore) and constrains is $0 < -s < 1, 1 - c < -s, 0 < 1 - c < -s < 1$ The final stable stage game is  is group 2 ($Type\ IB$) in two person game ($\begin{bmatrix} (1,1-c) & (-s,-s) \\ (0,-c) & (0,0) \end{bmatrix}$ is the original payoff matrix for avoidant game. the constrains of group 2 is $0 < -s < 1, 1 - c > -s$) The state transitions: $Type\ IIA \rightarrow Type\ IA2 \rightarrow Type\ IB$ The final evolving matrix NE: secure attachment.



Case 6: r = 3/2 (r=2 is the same state transition)

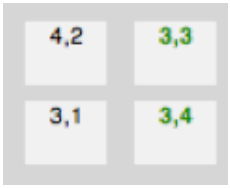
The final evolving matrix NE: secure attachment as well, similar with case 5.

The only difference is that it occurs two types state in transition which are not defined in groups of two person game.



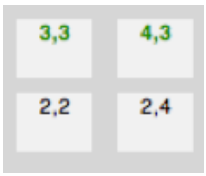
is very close to Type IA1 but it violate the constrains of $0 < -s$.

($\begin{bmatrix} (1,1-c) & (-s,-s) \\ (0,-c) & (0,0) \end{bmatrix}$ is the original payoff matrix for two person game.)

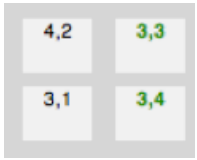


has pure nash equilibrium of both (Go, Ignore), (Don't Go, Ignore).

(Go, Ignore) $\epsilon\ \eta$ which is $\{\{(\mathbf{Go}, \mathbf{Attend}), (\mathbf{Go}, \mathbf{Ignore})\}\}$ for child.



is the state before reach to secure state.

The part of reinforcement rule :(Go Attend) occupied more shares after 

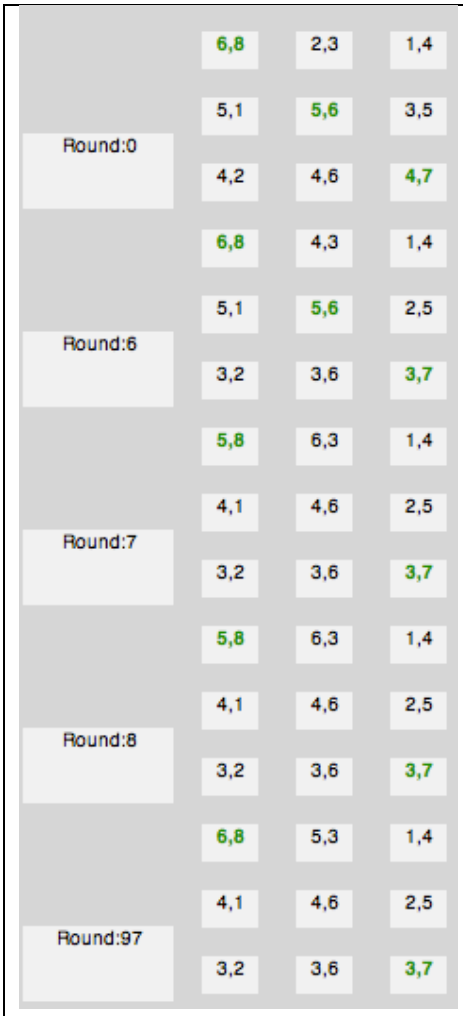
.

Figure 4.3 evolving ordinal payoff matrix of avoidant game

Avoidant Attachment for reinforcement rule are $\eta = \{\{(\mathbf{Go}, \mathbf{Attend}), (\mathbf{Go}, \mathbf{Ignore})\}\}$ for agents of game

Hostile Mother with $\eta = \{(\mathbf{Go}, \mathbf{Attend}), (\mathbf{Don't\ Go}, \mathbf{Attend}), (\mathbf{Freeze}, \mathbf{Attend})\}$ for parent, $\eta = \{(\mathbf{Go}, \mathbf{Attend}), (\mathbf{Go}, \mathbf{Ignore}), (\mathbf{Go}, \mathbf{Frighten})\}$ for child.

Group1 Games	*Type*				NE:
$0 - u < 1 - f - u$	*type ABCDEFGHIJKLMNO*	$(\mathbf{6}^*, \mathbf{8}^*)$	(2,3)	(1,4)	(Go, Attend) &
$1 - u < 1 - f - u$		(5,1)	$(\mathbf{5}^*, \mathbf{6}^*)$	(3,5)	(Don't Go, Ignore)&
$0 < 1 - f - u$		(4,2)	(4,6)	$(\mathbf{4}^*, \mathbf{7}^*)$	(Freeze, Frighten)
$-g - f < 1 - f - u$					



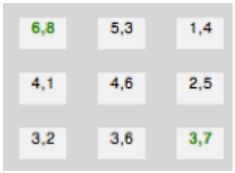
Case 7: $1 < r < 3/2$ (r = 1.2)

For Type $ABCDEFGHIJKLMnoPQr$ games, the parent's initial payoff matrix (and therefore the starting state for the game) is $\begin{bmatrix} 8 & 3 & 4 \\ 1 & 6 & 5 \\ 2 & 6 & 7 \end{bmatrix}$, the final state is also $\begin{bmatrix} 8 & 3 & 4 \\ 1 & 6 & 5 \\ 2 & 6 & 7 \end{bmatrix}$.

There is no state transition in parent evolving matrix.

The child's initial payoff matrix (and therefore the starting state for the game) is $\begin{bmatrix} 6 & 2 & 1 \\ 5 & 5 & 3 \\ 4 & 4 & 4 \end{bmatrix}$, the final state is $\begin{bmatrix} 6 & 5 & 1 \\ 4 & 4 & 2 \\ 3 & 3 & 3 \end{bmatrix}$.

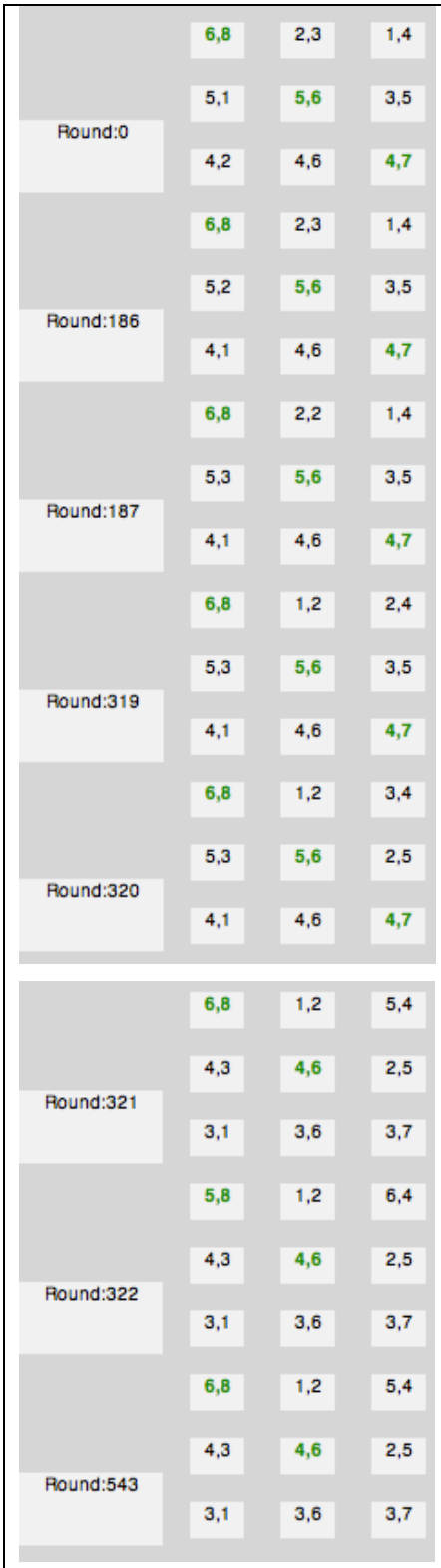
The final stable stage matrix:



the equilibrium(6*, 8*) corresponds to secure attachment and (3*, 7*) disorganized attachment

The final evolving matrix NE: secure and avoidant attachment.

Secure attachment also the pareto optimality



Case 8: r = 3/2 (r =2 is the same state transition)

the parent's initial payoff matrix (and therefore the starting state for the game) is $\begin{bmatrix} 8 & 3 & 4 \\ 1 & 6 & 5 \\ 2 & 6 & 7 \end{bmatrix}$, the final state is also $\begin{bmatrix} 8 & 2 & 4 \\ 3 & 6 & 5 \\ 1 & 6 & 7 \end{bmatrix}$.

The child's initial payoff matrix (and therefore the starting state for the game) is $\begin{bmatrix} 6 & 2 & 1 \\ 5 & 5 & 3 \\ 4 & 4 & 4 \end{bmatrix}$, the final state is $\begin{bmatrix} 6 & 1 & 5 \\ 4 & 4 & 2 \\ 3 & 3 & 3 \end{bmatrix}$.

The final stable stage matrix:

6,8	1,2	5,4
4,3	4,6	2,5
3,1	3,6	3,7

the equilibrium(6*, 8*) corresponds to secure attachment and (4*, 6*) corresponds to disorganized attachment which is classified as group 2

$$\begin{matrix} (1,1) & (-s,-s) & (-g,-g-f) \\ [\ (0,0-u) & (0,0) & (-i,-i-f-u)] \\ (-d,1-u) & (-d,0) & (-d,1-f-u) \end{matrix}$$

The final evolving matrix NE: secure and disorganized attachment.

Secure attachment also the pareto optimality

Figure 4.4 evolving ordinal payoff matrix of 'hostile' mother game

$\eta = \{$(Go, Attend), (Don't Go, Attend), Freeze, Attend$\}$ for parent of game

$\eta = \{$(Go, Attend), (Go, Ignore), (Go, Frighten)$\}\}$ for child of game

Affective Communication Error with $\eta = \{(\mathbf{Go}, \mathbf{Attend})\}$ for both child and parent.

Group2 Games $h - c < -s$ $0 - c < -m - a$ $0 - a > -t$ $-m - a > -s$	$Type$ $\boldsymbol{ABCDEFGHIJKLMnopQr}$	(6,3) (5,2) (4,1)	(2,5) $(\mathbf{4}^*, \mathbf{7}^*)$ (4,7)	(1,4) (3,6) $(\mathbf{4}^*, \mathbf{8}^*)$	NE: (Half Go, Half Attend) : disorganized attachemnt & (Don't Go, Ignore): avoidant attachment

	6,3	2,5	1,4
	5,2	4,7	3,6
Round:0	4,1	4,7	4,8
	6,4	2,5	1,3
	5,2	4,7	3,6
Round:3	4,1	4,7	4,8
	6,5	2,4	1,3
	5,2	4,7	3,6
Round:4	4,1	4,7	4,8
	6,6	2,4	1,3
	5,2	4,7	3,5
Round:5	4,1	4,7	4,8
	6,7	2,4	1,3
	5,2	4,6	3,5
Round:10	4,1	4,6	4,8
	6,8	2,4	1,3
	5,2	4,6	3,5
Round:12	4,1	4,6	4,7

Case 9: $1< r < 3/2$ (r = 1.2)

For ***type ABCDEFGHIJKLMnopQr*** games, the parent's initial payoff matrix (and therefore the starting state for the game) is $\begin{bmatrix} 3 & 5 & 4 \\ 2 & 7 & 6 \\ 1 & 7 & 8 \end{bmatrix}$, the final state is $\begin{bmatrix} 8 & 4 & 3 \\ 2 & 6 & 5 \\ 1 & 6 & 7 \end{bmatrix}$.

The child's initial payoff matrix is $\begin{bmatrix} 6 & 2 & 1 \\ 5 & 4 & 3 \\ 4 & 4 & 4 \end{bmatrix}$, the final state is $\begin{bmatrix} 6 & 2 & 1 \\ 5 & 4 & 3 \\ 4 & 4 & 4 \end{bmatrix}$.

The final stage game is

6,8	2,4	1,3
5,2	4,6	3,5
4,1	4,6	4,7

the equilibrium(6*, 8*) corresponds to secure attachment and (4*,6*) correspond to disorganised attachment, (4*,7*) correspond to avoidant attachment which is in Group 4

（The original stage game:

$$\begin{matrix} (1,1-c) & (-m,-m-a) & (-s,-s) \\ [(h,h-c) & (0,0-a) & (-t,-t)] \\ (0,0-c) & (0,0-a) & (0,0) \end{matrix}$$

Group 4's constrains as below:

`a' branch: $\mathbf{1-c>-m-a}$

`c' branch: $\mathbf{1-c>-s}$

` G ' branch: $\mathbf{h-c<0-a}$

`H' branch: $\mathbf{h-c<-s}$

`J' branch: $\mathbf{h-c<0}$

`n' branch: $\mathbf{0-a>-s}$

`o' branch: $\mathbf{0-a>-t}$)

The final evolving matrix NE: secure attachment and disorganized, avoidant attachment. Secure attachment also the pareto optimality

Round:0 6,3 2,5 1,4 5,2 4,7 3,6 4,1 4,7 4,8 Round:4239 6,4 2,5 1,3 5,2 4,7 3,6 4,1 4,7 4,8 Round:4292 6,6 2,4 1,3 5,2 4,7 3,5 4,1 4,7 4,8 Round:12764 6,8 2,4 1,3 5,2 4,6 3,5 4,1 4,6 4,7	**Case 10: $r = 3/2$ (same with r> 3/ 2)** Similar with case 9 The final evolving matrix NE: secure attachment and disorganized, avoidant attachment. Secure attachment also the pareto optimality

Figure 4.5 evolving ordinal payoff matrix of ‘affective communication error’ game using reinforcement rule$\{(\mathbf{Go}, \mathbf{Attend})\}$

Affective Communication Error with $\eta = \{(\mathbf{Go}, \mathbf{Attend}), (\mathbf{Half\ Go}, \mathbf{Attend}), (\mathbf{Don't\ Go}, \mathbf{Attend})\}$ for parent, and $\eta = \{(\mathbf{Go}, \mathbf{Attend}), (\mathbf{Go}, \mathbf{Half\ Attend}), (\mathbf{Go}, \mathbf{Ignore})\}$ for child.

Round:0	6,3	2,5	1,4
	5,2	4,7	3,6
	4,1	4,7	4,8
Round:6	6,3	3,5	1,4
	5,2	4,7	2,6
	4,1	4,7	4,8
Round:7	6,3	4,5	1,4
	5,2	3,7	2,6
	3,1	3,7	3,8
Round:9	6,3	5,5	1,4
	4,2	3,7	2,6
	3,1	3,7	3,8
Round:10	5,3	6,5	1,4
	4,2	3,7	2,6
	3,1	3,7	3,8

Round:179	5,4	6,5	1,3
	4,2	3,7	2,6
	3,1	3,7	3,8
Round:180	5,5	6,4	1,3
	4,2	3,7	2,6
	3,1	3,7	3,8
Round:212	5,6	6,4	1,3
	4,2	3,7	2,5
	3,1	3,7	3,8
Round:213	5,7	6,4	1,3
	4,2	3,6	2,5
	3,1	3,6	3,8
Round:227	5,8	6,4	1,3
	4,2	3,6	2,5
	3,1	3,6	3,7
Round:546	6,8	5,4	1,3
	4,2	3,6	2,5
	3,1	3,6	3,7

Case 11: $1< r < 3/2$ (r = 1.2)

For ***type ABCDEFGHIJKLMnopQr*** games, the parent's initial payoff matrix (and therefore the starting state for the game) is $\begin{bmatrix}3 & 5 & 4\\2 & 7 & 6\\1 & 7 & 8\end{bmatrix}$, the final state is $\begin{bmatrix}8 & 4 & 3\\2 & 6 & 5\\1 & 6 & 7\end{bmatrix}$.

The child's initial payoff matrix is $\begin{bmatrix}6 & 2 & 1\\5 & 4 & 3\\4 & 4 & 4\end{bmatrix}$, the final state is $\begin{bmatrix}6 & 5 & 1\\4 & 3 & 2\\3 & 3 & 3\end{bmatrix}$.

The intermit state at round 7 is

6,3	4,5	1,4
5,2	3,7	2,6
3,1	3,7	3,8

The NE changes to

(Go, Half-Attend) and (Don't Go, Ignore)

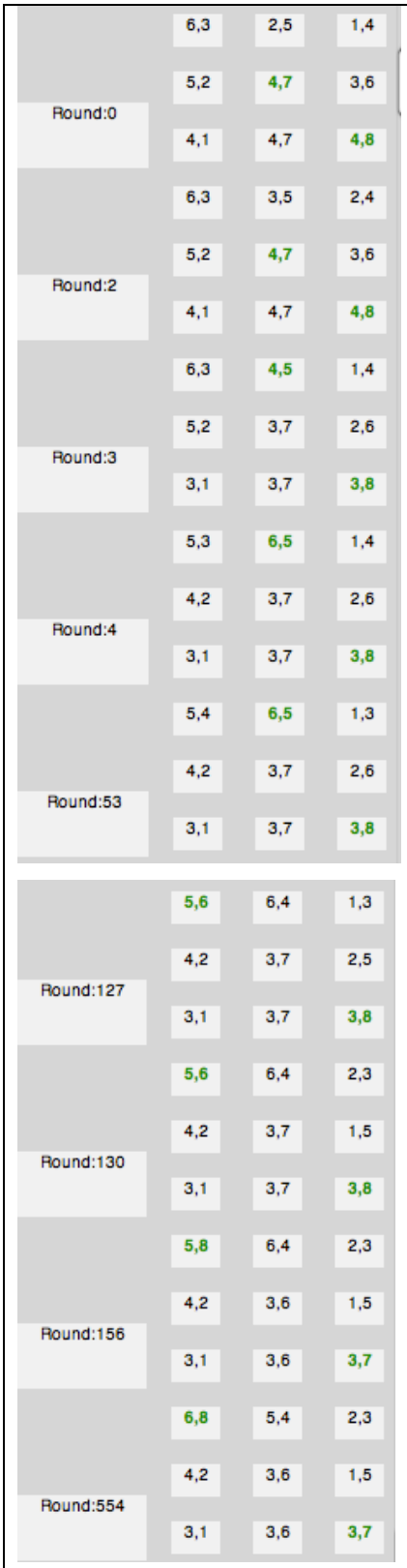
Though the intermit state, we can forecast the further state transition. Since (Go, Half-Attend) $\epsilon\, \eta$, it will also have certain effect on(Go, Attend)

The final stable stage game is

6,8	5,4	1,3
4,2	3,6	2,5
3,1	3,6	3,7

the equilibrium(6*, 8*) corresponds to secure attachment and (3*, 7*) avoidant attachment, which is close to the Group3 of 'affective communication error' games but violates the constrain: $\mathbf{h} - \mathbf{c} > 0 - \mathbf{a}$

The final evolving matrix NE: secure attachment and avoidant attachment. Secure attachment also the pareto optimality. The final evolving matrix NE: secure attachment



Case 12: r = 3/2

Similar with case 11 (r > 3/2 is also the same state transitions)

The final evolving matrix NE: secure attachment and avoidant attachment. Secure attachment also the pareto optimality

The final stage game :

6,8	5,4	2,3
4,2	3,6	1,5
3,1	3,6	3,7

Figure 4.6 evolving ordinal payoff matrix of 'affective communication error' game using reinforcement rule {(Go, Attend), (Don't Go, Attend), Freeze, Attend} for parent and reinforcement rule {(Go, Attend), (Go, Half Attend), (Go, Ignore)} for child

This section shows detailed evolving progress of state transition on avoidant attachment game (2X2 matrix games) and disorganised attachment game (3X3 matrix games). The results of those simulations have all reach in an expected final stable matrix or the goal matrix that involving secure attachment style, as to reinforcement rule is powerful and result-oriented.

4.3.2 Experimental Results

We set the experiment repeat every simulation independently 500 times to collect the average number of rounds before the interaction converged towards a stable, secure relationship. The experiment process was repeated for various combinations of the parameters for the discounted factor δ, and reinforcement rate. For example, r = 1.01, r = 1.02 and r = 1.03, and for the reinforcement rules for parent in avoidant attachment, η = {(Go, Attend), (Don't Go, Attend)} and η = {(Go, Attend)}, for child, η = {(Go, Attend), (Go, Ignore)} and η = {(Go, Attend)}. In all games we set the parent's exploration parameter to k = 2.

In the charts below we have plotted series of various reinforcement rates r against the varying discount factors δ on the horizontal axis, and the average number of rounds required before stable, secure attachment emerged on the vertical axis. A solid line indicates that the reinforcement rule $\eta = \{(\text{Go}, \text{Attend})\}$ was used for all three attachments, and a dashed line that the reinforcement rule $\eta = \{(\text{Go}, \text{Attend}), (\text{Don't Go}, \text{Attend})\}$ was used for parent in avoidant attachment, and $\eta = \{(\text{Go}, \text{Attend}), (\text{Don't Go}, \text{Attend}), (\text{Freeze}, \text{Attend})\}$ for parent in "hostile" mother pareting profile. $\eta = \{(\text{Go}, \text{Attend}), (\text{Half Go}, \text{Attend}), (\text{Don't Go}, \text{Attend})\}$ also for parent in "Affective Communication Error" parenting profile, there are the corresponding reinforcement for child as well.

Avoidant attahcment for different reinforcement rules are for both sides of game

Group3 Games $0 > -s$ & $1 - c < -s$	*Type IIA*:	$\begin{bmatrix}(4,2) & (2,3)\\(3,1) & (3^*,4^*)\end{bmatrix}$	NE: *avoidant* (Don't Go, Ignore)

Fisrtly the experiments were run from particular initial game ($Type\ IIA$), and the average number of games required before a stable, secure attachment relationship emerged is shown in figure 4.7.

A similar observation with Cittern was that dyads had a large discount factor $\delta > 0.7$ (i.e. both sides placed more of a preference on future rewards) did not on average converge to a stable, secure attachment style within 10,000 rounds of play, and we can explain in algorithm and rules chosen. If the action-combination outcome from some arbitrary round $(a_c, a_p) \in \eta$ then the reward issued is based on a reinforcement of the parent's underlying payoff matrix (such that rewards are

potentially unbounded). The parent will receive a reward even when no state transition occurs, i.e. R_a (s,s′) > 0 for s = s′. Since state transitions only occur when these reinforcements result in a new ordinal payoff matrix for the parent, we can have the situation whereby Q values for Q(s, a) are updated even if no state transition has occurred. A parent with a high discount factor δ > 0.6 who initially selects actions in state s that do not result in any state transitions will learn relatively larger Q value updates for these non-reinforceable actions than would a parent with a lower discount factor. As these Q values are successively updated and grow larger, the probability of selecting actions in state s associated with non-optimal Q values becomes smaller, and it appears as though there is a threshold for δ for which the initial non-transitional Q value updates are so large that the probability of exploration becomes too small to result in a change in attachment style. Therefore, to see a convergence toward secure attachment within the number of rounds we have considered and in parents with a discount factor of δ > 0.7, we postulate that a lower value of k would be required, resulting in more state-action exploration.

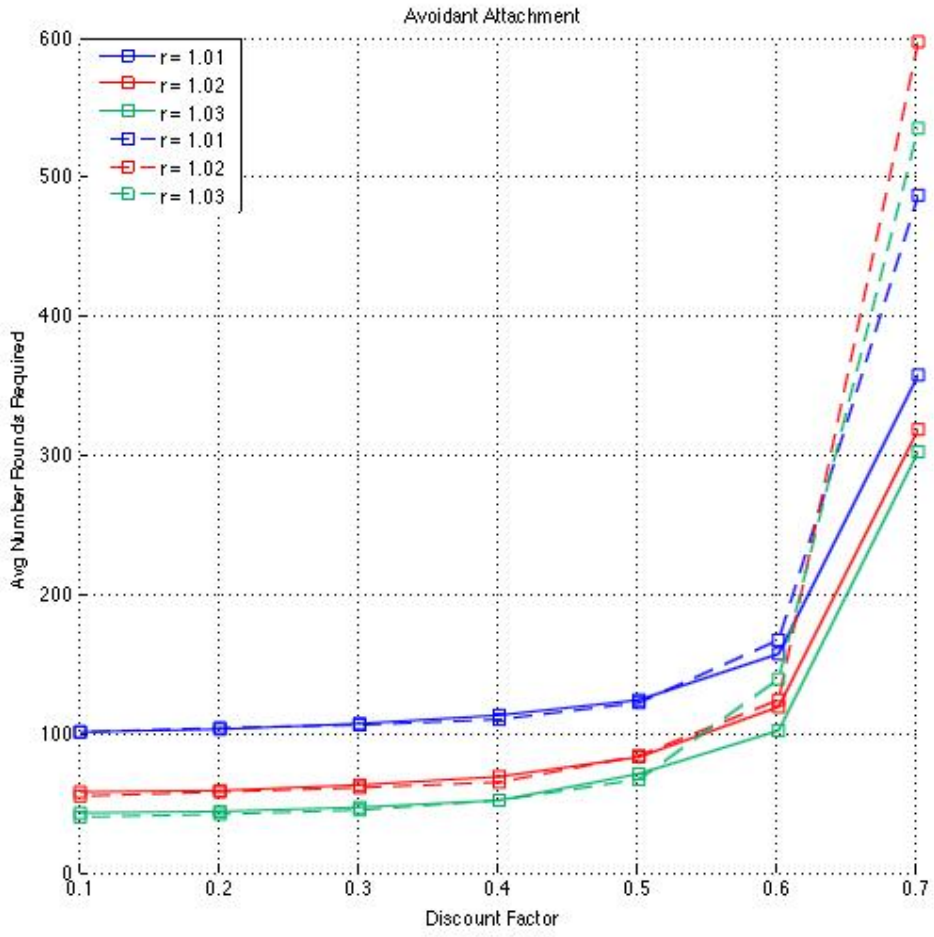


Figure 4.7: The average number of rounds required for a stable, secure attachment style to evolve when both parent and child played Q-learning from avoidant Attachment *Type IIA*. For various discount factors (δ) and reinforcement rates (r). Solid lines are results for the reinforcement rule

for parent and child both are $\eta = \{(\text{Go}, \text{Attend})\}$ and dashed lines for parent is $\eta = \{(\text{Go}, \text{Attend}), (\text{Don't Go}, \text{Attend})\}$and for child is $\eta = \{(\text{Go}, \text{Attend}), (\text{Go}, \text{Ignore})\}$.

Compared to Cittern's experiment for single-agent, only parent played Q-learning in game beginning from $Type\ IIA$, not matter which best response strategies for child using, BRTLM or MEP, the multi-agent Q-learning have about 5-fold better performance than single-agent Q-learning.

Group2 Games $h - c < -s$ $0 - c < -m - a$ $0 - \text{a} > -\text{t}$ $-m - a > -s$	$Type$ $\boldsymbol{ABCDEFGHIJKLMnopQr}$	(6,3) (2,5) (1,4) (5,2) $(\mathbf{4}^*, \mathbf{7}^*)$ (3,6) (4,1) (4,7) $(\mathbf{4}^*, \mathbf{8}^*)$	NE: (Half Go, Half Attend) : disorganized attachemnt & (Don't Go, Ignore): avoidant attachment

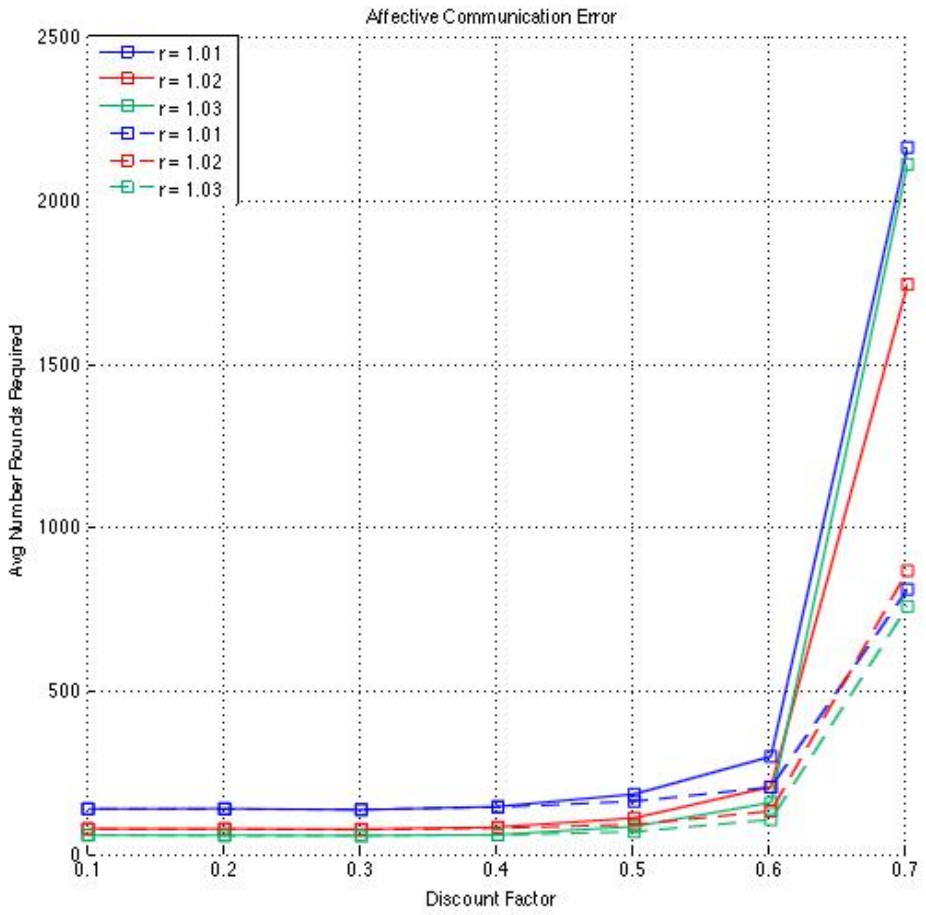


Figure 4.8: The average number of rounds required for a stable, secure attachment style to evolve when both parent and child played Q-learning from Affective Communicat Attachment $Group$ 2 ($type\ ABCDEFGHIJKLMnopQr$).

For various discount factors (δ) and reinforcement rates (r). Solid lines are results for the reinforcement rule for parent and child both are $\eta = \{(\text{Go}, \text{Attend})\}$and dashed lines for parent is $\eta = \{(\text{Go}, \text{Attend}), (\text{Half Go}, \text{Attend}), (\text{Don't Go}, \text{Attend})\}$ and for child is $\eta = \{(\text{Go}, \text{Attend}), (\text{Go}, \text{Half Attend}), (\text{Go}, \text{Ignore})\}$.

Group1 Games $0-u<1-f-u$ $1-u<1-f-u$ $0<1-f-u$ $-g-f<1-f-u$	$Type$ $type\ ABCDEFGHIJKLMNO$	$(\mathbf{6}^*,\mathbf{8}^*)$ $(\mathbf{5},\mathbf{1})$ $(\mathbf{4},\mathbf{2})$	$(\mathbf{2},\mathbf{3})$ $(\mathbf{5}^*,\mathbf{6}^*)$ $(\mathbf{4},\mathbf{6})$	$(\mathbf{1},\mathbf{4})$ $(\mathbf{3},\mathbf{5})$ $(\mathbf{4}^*,\mathbf{7}^*)$	NE: (Go, Attend) & (Don't Go, Ignore)& (Freeze, Frighten)

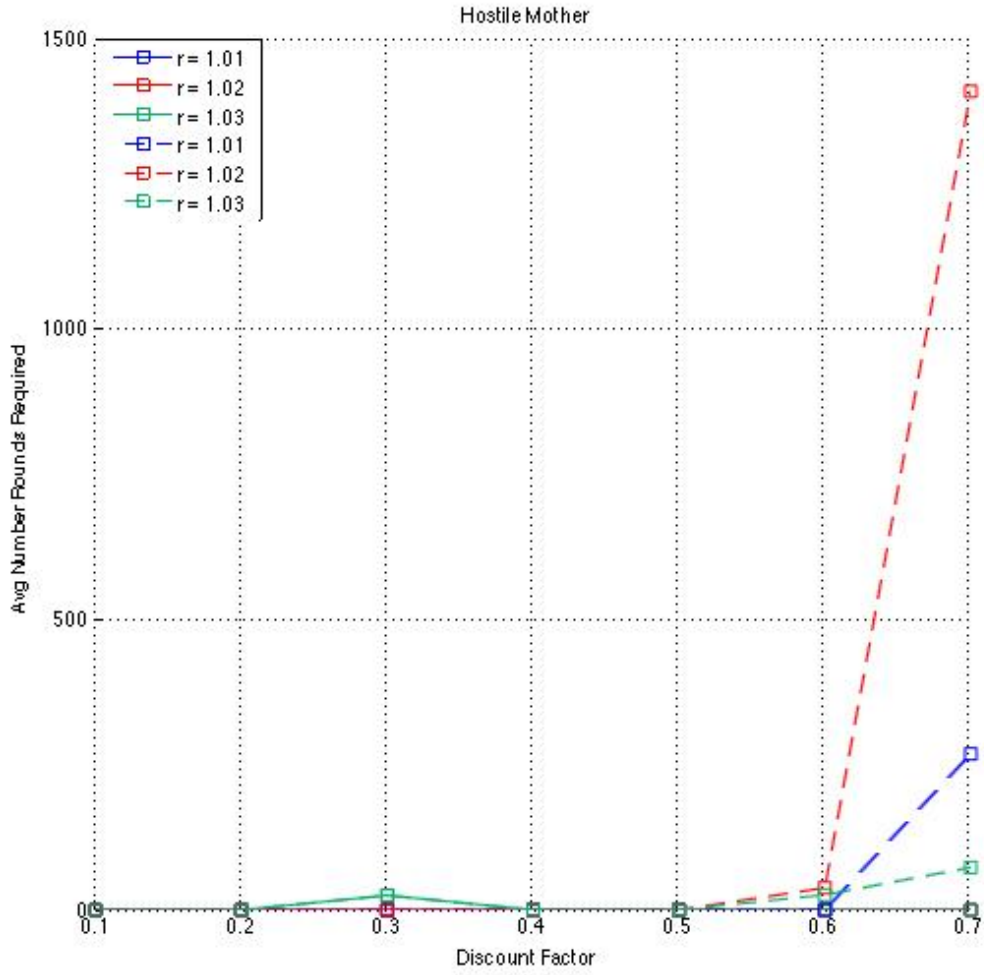


Figure 4.9: The average number of rounds required for a stable, secure attachment style to evolve when both parent and child played Q-learning from model of disorganised attachment : 'Hostile' Mother $Group\ 1(type\ ABCDEFGHIJKLMNOP)$. For various discount factors (δ) and reinforcement rates (r).

Solid lines are results for the reinforcement rule for parent and child both are $\eta = \{(\text{Go}, \text{Attend})\}$ and dashed lines for parent is $\eta = \{(\text{Go}, \text{Attend}), (\text{Don't Go}, \text{Attend}), (\text{Freeze}, \text{Attend})\}$ and dashed lines for parent is and for child is $\eta = \{(\text{Go}, \text{Attend}), (\text{Go}, \text{Ignore}), (\text{Go}, \text{Frighten})\}$ respectively.

When consider 'Hostile' Mother model, since in each of these games the secure Nash equilibrium is Pareto optimal, furthermore, when $\eta = \{(\text{Go}, \text{Attend}), (\text{Don't Go}, \text{Attend}), (\text{Freeze}, \text{Attend})\}$ is used, there are more possible transitions out of each state, resulting in a far more complicated state traversal. The solid lines graph looks different from 'Avoidant' two person game, and 'Affective Communicaton Error' because the secure attachment is already one of the original games, after the $\eta = \{(\text{Go}, \text{Attend})\}$ reinforced, the converge of secure stable state costs few rounds to get.

CHAPTER 5

Evaluation

The success of the project is measured by whether or not it has satisfied the characteristics of game theoretic model of Strange Situation, and can observe the process evolution of matrix game from our simulation and experiments.

5.1 Strengths of the project

The major strength with my project is that programming of simulation obeys the project setting and algorithm. It is very easy to use. During the single iteration simulation phase, game system shows exhaustive game state transitions and solving information which lists entire game's variables about choices of agents and also the variables about state transition, which can show how the strategy improvement algorithm works. Furthermore a graph showed the choice of each round is easy to observe the trend of the game in a holistic scale. And the experiment is easy to use as well. When do the experiment, only change 2 to 3 parameters value, such as reinforcement rule or discount factor or reinforcement rate. The system can output all the data and graph it can be used to analysed the results.

Strength of the project has been my ability to focus the scope of the project so as to allow myself to complete the aims that I had decided upon during the specification phase. I managed to meet all of the aims for the project that is strength of the project in itself.

5.2 Weakness of the project

One of the aspects of the project that I feel has been a weakness is that although the game system can run smoothly and output correct and reliable data for experiment use, the algorithm we adopt the simple Q-learning algorithm David Cittern used in his project. When simple Q-learning used in multi-agent games, it have not fully used the stochastic game framework, which could have not considering the joint action and states in transition function $S \times A \times S \rightarrow [0,1]$. The experiment strategy we use is 'Boltzmann' action selection rule, which is not only an optimal policy instructing the agent to always choose the action gained highest expected reward, but also not stuck in local maximum. The exploration vs exploitation is abstract concept, and the further experimental verification can explore this part, which we did limited work on it. So far we did was that, under our model, we observed that parents with large discount factors $\delta > 0.6$ (i.e. parents who placed a relatively large importance on future rewards over current rewards) hampered the evolution of the

attachment style at their dyad from avoidant to secure, and postulated that this was due to the exploration parameter *k*.

The chapter 3 gives us more inspiration on classical and updated algorithms used in multi-agent games. Although we did not apply those algorithms in this project, but it extended the knowledge of learning mechanism, such as that Hu[19] extend Q-learning to a non-cooperative multi-agent context (grid game) framework of general-sum stochastic games, called Nash Q learning. Nash-Q generalizes single-agent Q-learning to multi-agent environments by updating its Q-function based on the presumption that agents choose Nash-equilibrium actions. Given some highly restrictive assumptions on the form of stage games during learning, the method is guaranteed to converge. Empirical evaluation on a pair of small but interesting grid games shows that the method can often find equilibria despite the violations of some theoretical assumptions.

The Q-learning used in our project, it satisfies basic parts of the standard definition of stochastic games, whereas in the action-chosen phase, which commonly uses a state transitions function, both agent's next actions should be concerned, but as to the experimental strategies we use, it simply let each agent learn their own state and only considering the own possible actions in next rounds. The both state be considering is realizable just using the same rules and algorithms of our project, whereas joint actions chosen is not suitable for the simple Q-learning plus 'Boltzman' action selection rule. The Nash-Q learning for general sum game is a good reference about our further work. The nash equilibrium finding in every stage game's nash-Q table is also reasonable to used in our project. If we can solve the setting of state-transition function which can obeys the standard stochastic game framework and also can obey the parent-child interaction game setting, Nash-Q learning algorithms is a good way to inspire our further work.

CHAPTER 6

Conclusions and Future Work

6.1 Conclusions

Firstly, we have reviewed the psychological theory of attachment in children including its origins, the strange situation protocol, the four classification types (secure, avoidant, ambivalent and disorganised), experimental observations regarding the stability of these classifications and arguments for causality. We have also summarised relevant ideas from game theory, which is the mathematical tool we will use to understand strange situation-type interactions, and in particular the concept of a Nash equilibrium which we will use to discuss attachment classifications and stability. Finally, we gave a detailed overview of a two-person, two-action game theoretic model of the strange situation that was proposed in another paper, which is used as a basis for this work and will be referred to frequently. In that paper, ordinal games were categorised according to the position of their Nash equilibria, and the authors have argued that these equilibria are indicative of secure, avoidant and ambivalent attachment types. David Cittern [1] introduced two models with the intention of capturing various aspects of disorganised attachment. The first is based on Lyons-Ruth's profile of a hostile parent, for which we introduced the `Frighten' action as an example of a negative-intrusive behaviour, and also attempted to incorporate an element of role-reversal within the parent's payoff structure. For the child, disorganisation was captured in the form of a `Freeze' action, representing the dissociation, which has been observed for disorganised children during the strange situation. For this model we identified specific parameter configurations resulting in games with a pure Nash equilibrium at (Freeze, Frighten), representative of a form of disorganised attachment, and also the mixed strategies. As to knowing various aspects of disorganised attachment, the two models (‘hostile’ mother and ‘affective communication Error’) he introduced were becoming the most important direct topics in this project. We consider an iterative game environment or how attachment classifications may come to change over repeated interactions have not completed in Cittern’s project. We feel that these are important areas of investigation and we attempt to tackle them in this report, beginning with the simulations of models of disorganised attachment in the iteration game chapter.

Besides the models we use, we consider that the child is great learner; we attempt to follow the Cittern’s previous work and extend his only-parent-learning framework to stochastic game for multi-agent games.

The remainding of the project has been concerned with how attachment styles could change over the course of repeated interactions between the parent and the child. In particular we have focused

on modeling scenarios in which the parent's payoff matrix evolves such that new pure Nash equilibria emerge. We began by looking at the iterated game of a single parent-child dyad, and assumed that there was some external force such as a psychotherapist encouraging and praising only certain outcomes to rounds of the iterated game. We combined this with a model of learning for the parent, whereby at each iteration of the game the parent re-evaluates the most optimal behaviour with regards to their current underlying 'emotional' state. Although based on a large number of assumptions on the part of the parent (such as a fixed exploration parameter and a decreasing learning rate), we have demonstrated how a combination of controlled outcome reinforcement and k values can result in the emergence of a stable, secure attachment style. We have also shown how a parent who places too much relative importance on future rewards may actually hinder rather than help this evolution. Perhaps most crucially, we have highlighted that the child need not be aware of any of these reinforcement or learning mechanisms in order for this desirable outcome to occur.

6.2 Future Work

When simulations, we have found that the existing model groups is incomplete of classification, thus, the further work can concern the development of detailed model groups are given out. David Cittern's disorganised models provide the basic framework of groups, it still have potential to observe more accurate and detailed classifications among various types. We know the psychology aspect is based on large number of parent- child experiments or observation on daily life, even in long run. Although lack of reality data of psychological experiment, we still contributes a inspiring view to investigate psychology using mathematical mechanism in which attachment relationships can be modeled, and that it provide more space for interdisciplinary research in this area. In order to achieve the real interdisciplinary research, the combination of reality experiment and model experiment should mutually support and avoid the mathematical assumptions get far beyond or deviate the original facts, such as the range of discount learning rate or exploration parameter k.

Another aspects we can follow the result oriented method to implement experiments. We already used reinforcement rules guide parent-child interaction to secure attachment style, the evolving progress shows a comparatively random state transitions. The ideal thought is find an exact transition function to implement the experiment, which can follow the rule of change in reality. This work is not impracticable in this project for a lack of reality data.

The last one but not the least one is we can attempt to apply newer multi-agent algorithms that mentioned in section 3.5. There is our limitation of this project not using more complex algorithms and also as to the simple algorithm we use, which to some extend not to perfectly obey the stochastic game framework. Another algorithms can be more feasible, such as Nash-Q learning using Nash Equilibrium to find goal states, or WoLF Policy Hill-Climbing to follow the "learning quickly while losing, slowly while winning'' purpose to converge to a expected stable state.

Appendices

Implementation Notes

The repeated game were implemented in Java. A front-end was created with Matlab so that parameters could be supplied for successive simulations without re-compilation, and for easy analysis of data results. The Iterated game of avoidant attachment can be run using the stability_analysis.m script, the affectivceCE can be run using the aff_analysis.m script, the hostileMother can be run using hostile_analysis.m script.

Bibliography

[1] D. Cittern. Models of Child-Parent Interaction in Game Theory. MSc dissertation, Computer Science Department of Imperial College London, 2011.

[2] M. Bowling. and M. Veloso. An Analysis of Stochastic Game Theory for Multiagent Reinforcement Learning. School of Computer Science Carnegie Mellon University, 2000.

[3] M. Bowling. and M. Veloso. Rational and Convergent Learning in Stochastic Games. School of Computer Science Carnegie Mellon University.

[4] D. Petters. Building agents to understand infant attachment behaviour. School of Computer Science, University of Birmingham.

[5] A. Amengual. A computational model of attachment secure responses in the Strange Situation. International Computer Science Institute, 2009.

[6] M. Ainsworth, M. Blehar, E. Waters, and S. Wall. *Patterns of Attachment: a psychological study of the strange situation*. Erlbaum, Hillsdale, NJ, 1978.

[7] J. Bowlby. *Attachment and loss: volume 1 attachment*. Basic books, New York, 1969-1982.

[8] L. S. Shapley. Stochastic games. Proceedings of the National Academy of Sciences of the United States of America, 39(10):1095–1100, 1953.

[9] J. Bowlby.and M. Ainsworth. Child Care and the Growth of Love. Pelican, 1953.

[10] M. Main. and J. Solomon. Procedures for identifying infants as disorganised/disoriented during the Ainsworth Strange Situation (In Attachment in the Pre-school Years). Uni- versity of Chicago Press, 1990.

[11] E. Waters., S. Merrick., D. Treboux., J.Crowell., and L. Albersheim. Attachment security in infancy and early adulthood: A twenty-year longitudinal study. Child Development, 71, 2000.

[12] L. A. Sroufe., B. Egeland., E. A. Carlson., and W. A. Collins.. The Development of the Person: The Minnesota Study of Risk and Adaptation From Birth to Adulthood. Guilford Press, 2005.

[13] P.Fonagy. Thinking about thinking: some clinical and theoretical considerations in the treatment of a borderline patient. International Journal of Psychoanalysis, 72:639-656, 1991.

[14] P. Fonagy., M. Steele., and H. Steele. Maternal representations of attachment predict the organisation of infant mother-attachment at one year of age. Child Development, 62 (5):891 - 905, 1991.

[15] D. Iwaniec .The emotionally abused and neglected child. John Wiley & Sons Ltd, 2006.

[16] M. J. Osborne., and A. Rubinstein. *A Course in Game Theory*. The MIT Press, 1994.

[17] R.A. Howard. *Dynamic Programming and Markov Processes*. MIT Press, 1960.

[18] M.L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, pages 157–163. Morgan Kaufman, 1994.

[19] Junling Hu and Michael P. Wellman. Multiagent reinforcement learning: Theoretical framework and an algo- rithm. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 242–250, San Francisco, 1998. Morgan Kaufman.

[20] M. Bowling. Convergence problems of general-sum multiagent reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 89–94, Stanford University, June 2000. Morgan Kaufman.

[21] C.J.C.H. Watkins. Learning from delayed rewards. PhD thesis, University of Cambridge, 1989.

[22] Myers J. The APSAC Handbook on Child Maltreatment. Sage Publications, 2002.

[23] D. Wallin. Attachment in Psychotherapy. Guilford Press, 2007.

[24] J Junling Hu and Michael P. Wellman. Nash Q-Learning for General-Sum Stochastic Games. Journal of Machine Learning Research 4,1039-1069, 2003.