Modelling Emotions and Simulating their Effects on Social Interactions in Agent Systems

Thesis submitted in accordance with the requirements of
the University of Liverpool for the degree of Doctor in Philosophy by

Martyn Lloyd-Kelly

March 2014

For my father, Philip Lloyd
24th January 1967 - 10th November 1989

Abstract

Agent-based decision-making usually relies upon game theoretic principles that are "rational" i.e. decision-making is purely mathematical based on utilities such as the wealth of an agent. In the context of public goods games, such reasoning can often lead to non-optimal, destructive outcomes for both individuals and the total system, as shown in many scenarios from game theory. This thesis considers how the use of *emotions* can impact upon decision-making and social interactions amongst agents in the iterated Prisoner's Dilemma game by modelling emotions in a functional manner.

The background to the thesis is first presented in chapters 2 and 3 where the argument for emotions being included in agent-based decision-making, and evidence to support this proposition, is outlined. Various philosophical issues are also considered such as: do emotions directly motivate an agent's intentional behaviour and, is an agent's decision-making still rational if emotions are used? The framework developed to allow for modelling of emotions in agents is then discussed in chapters 4 and 5 where major psychological models of emotion and computational implementations thereof are discussed. Finally chapters 6 to 8 present extensive investigations into how the emotions modelled using the framework affect social interactions amongst agents in the context described above. As of yet, this topic has been relatively unexplored by computer science and there is space for novel, interesting contributions to be made, these contributions are outlined below.

In chapter 6 the emotions of *anger* and *gratitude* are modelled and their effects upon social interactions are analysed. In particular, I look at whether agents endowed with these emotions offer any improvement upon the success of agents using with the "tit-for-tat" strategy when playing against other leading strategies from Axelrod's famous computer tournament. How these emotions affect rates of cooperation/defection and the fairness of individual scores is considered along with why they do so.

This investigation is furthered in chapter 7 where *admiration* is modelled and an investigation is performed into what emotional characters are selected for under different initial conditions and why. This examination provides a discussion regarding what emotional social norms emerge in a population when agents admire the individual success of others. Two salient questions are asked: is it is the case that emotional characters which promote the total wealth of the system are selected for as an emergent property and, do different initial conditions affect the emotional characteristics selected for?.

Finally, chapter 8 extends chapter 7 by modelling *hope* and enquires as to how particular emotional character populations (after a complete social norm has been established) deal with destabilisation of cooperation cycles due to periodic defection. The performance of agents endowed with differing emotional characters are again tested under different initial conditions and specific behavioural features of particular emotional characters are considered. In doing this I comment upon how different emotional characters deal with periodic defection and what the best approach is both in context of an agent's individual score and the total score of the system.

Acknowledgements

"For we put the thought of all that we love into all that we make."

J.R.R. Tolkien

Selecting names to mention in this section has been quite a challenge so I extend my sincerest apologies if I have not mentioned certain people; if I have been remiss in text I have not been in thought.

First, I would like to thank my family for their immeasurable support, encouragement and love; your belief in me means more than I could ever hope to explain. In particular, I would like to acknowledge my uncle: Dominic. You have given me a great amount of help during the writing of this thesis (and prior to it) with no expectation of repayment. Without this aid I do not believe that I would have even come close to completing; thank you for being a father in his absence. I also wish to thank Jackie for providing me with somewhere stable when things got difficult at home, I hope this little mention lets you know how much I appreciate your kindness. Last but by no means least, I want to acknowledge my Mum for raising me with dignity, strength and unconditional love. Words would never be able to express my gratitude towards you.

I also wish to acknowledge my partner: Gemma. Since meeting you my life has taken a turn for the better and I do not think I could love anybody more than I do you. I realise that during the writing process of this thesis I have been inordinately pre-occupied with work but your acceptance and patience with regards to this, along with your constant encouragement and understanding, has been invaluable. I do not know if you fully appreciate everything you have done for me since we met but, if not, I hope that I can somehow convey how much of a source of strength, love and motivation you have been.

I must also acknowledge my computer-science office mates: Mr. (hopefully soon, Dr.) Anton Minnion, Mr. Luke Riley, Dr. Muhammad Khan and Dr. Adam Wyner. My time with you all was both enlightening and immensely humorous; G22 is sorely missed and will continue to be long after I have moved onto new ventures. Throughout my Ph.D. I was very lucky to have known such intelligent, friendly and funny people, I could not have wished for a better set of friends. I wish you all the very best in your future careers and I sincerely hope that we stay in touch. Anton: thank you for being a good friend, keep working hard; I'll see you on the other side. Muhammad: I wish you and your family all the very best I possibly can. Thank you for the interesting and thought-provoking discussions we shared. Luke: thank you for the laughs, collaborations and drinks. Keep eschewing that stereotype of the typical computer science Ph.D. student! Adam: I really wish we could have worked together more in your time at the university but, for the times we did, thank you. Your help with all aspects of my career has been

invaluable and I hope we'll meet up at some exotic conference once again and laugh as we always used to.

Also, I must thank the Engineering and Physical Sciences Research Council (EPSRC) of Great Britain for their financial support for the initial three years of my thesis. I would also like to thank the University of Liverpool, the department of Computer Science and the faculty of science for their provision of financial aid and training throughout my time as a Ph.D. student. Without this help I would not have been able to promote my research, meet others from around the world who share in my passion for science or have expanded my world-views in the way that I have.

I would also like to extend my deepest gratitude and utmost respect for my Ph.D. supervisors: Professor Trevor Bench-Capon, Dr. Katie Atkinson and Professor Peter McBurney. Your insights, belief and hard work throughout my time as a Ph.D. student have been immeasurable. Completing this work is rather bitter-sweet in that, I am glad to have finally finished but I am quite sad that we will no longer be able to regularly sit and converse about topics of interest. You have given me some of the most important academic skills imaginable: the ability to present arguments clearly, the ability to communicate to others about my work, the ability to write persuasively and the ability to undertake and produce quality research. Being supervised by each of you has been an absolute pleasure and I can only hope to become as well-respected and successful should I continue with a career in academia. Thank you for everything, I am forever indebted to each of you.

Finally, I would like to acknowledge two incredibly important people without whom I would not have become the man I am today: my grandmother, Joyce Kelly, who passed away during the writing of this thesis on the 25th of December 2012 and my father, Philip Lloyd, who passed away on the 10th of November 1989. Both of your deaths have left holes in my life that could never be filled. I love you both so very much and although I can never truly express the loss I feel for both of you, I hope that this small acknowledgement will do you both justice. Nan: your friendship, belief in me and your open invitation of a warm bed, food and a place to call home will always live with me. The aid you provided, with no expectation of repayment, is more appreciated than you could ever imagine and for that I thank you from the bottom of my heart. My only regret is that I did not finish this thesis in time for you to see its completion. Dad: above all else, I hope I have done you proud. Everything I have ever done I have done with the intention of honouring your memory. This may seem foolish since I have had a lot to honour especially since your family and friends hold you in the highest regard possible and the love expressed for you is truly awe inspiring. You are my greatest inspiration, my greatest motivation and my greatest loss.

Contents

Illustrations

List of Figures

List of Tables

Notations

The following notations and abbreviations are found throughout this thesis:

ACRES	**A**rtificial **C**oncern **R**ealisation **S**ystem
BDI	**B**eliefs-**D**esires-**I**ntentions
CC	**C**ooperate-**C**ooperate or "mutual cooperation" (in context of the Prionser's Dilemma)
CD	**D**efect-**C**ooperate (in context of the Prionser's Dilemma)
CPU	**C**entral **P**rocessing **U**nit
DC	**D**efect-**C**ooperate (in context of the Prisoner's Dilemma)
DD	**D**efect-**D**efect or "mutual defection" (in context of the Prisoner's Dilemma)
EBDI	**E**motional-**B**eliefs-**D**esire-**I**ntentions
GTFT	**G**enerous **T**it-**F**or-**T**at
MAS	**M**ulti **A**gent **S**ystem(s)
OCC	**O**rtony-**C**lore-**C**ollins
OCEAN	**O**penness-**C**onscientiousness-**E**xtraversion-**A**greeableness-**N**euroticism
TFT	**T**it-**F**or-**T**at
VAF	**V**alue-Based **A**rgumentation **F**ramework

Chapter 1

Introduction

"You don't get emotions by manipulating 0s and 1s."

John Searle

Homo economicus, or *the economic human*, has long been the archetypal model of agency in the context of computer-science and economics. The origin of the term is a research subject in itself with O'Boyle [138] ascribing the earliest known use of the term to Panteleoni [143] in 1889. The model casts humans as rational and self-interested; actors who base all decisions upon a potentially ruthless pursuit of maximisation of subjective utility. The use of *homo economicus* as a model of agency in computer science is especially apparent when modelling or discussing the decision-making of agents in the context of multi-agent systems (hereafter referred to as MAS). The problem however, is that the breadth of human behaviour observed since the dawn of *homo sapiens* cannot be solely explained by such a narrow philosophy. In fact, some researchers in the field of behavioural economics [50], [120] have rejected the notion of *homo economicus* and have turned to a consideration of the effects of *emotion* upon decision-making. In doing so, these researchers hope that the role of emotion in decision-making may be understood, modelled and thus taken into consideration when predicting human behaviour in an economic context.

Homo economicus casts human beings as being logically adept and entirely rational with respect to its consideration of actions. Indeed the rationality associated with *homo economicus* depends upon logical adeptness but *homo sapiens* are not always as logically adept as they would like to believe. Major psychological research such as [203], [71] and [100] highlights the enormous capacity human beings possess when it comes to demonstrating and engaging in irrational behaviour and these works are but a small selection of those available. Works such as those above also show that *homo sapiens* have a number of innate coping mechanisms for the incredibly complex environment they inhabit. These innate coping mechanisms motivate humans to focus attention

upon certain issues and ignore others, even though probabilistically and statistically-speaking, the issues ignored are just as relevant as they ever were. Gardner in [69] provides a comprehensive discussion of major logical flaws observed in humans and proposes that with respect to decision-making, *homo economicus* $\neq$ *homo sapiens*.

Determining what separates *homo economicus* from *homo sapiens* is an enormous task but something must, otherwise the observed dissonance between the two models of agency would not exist. Simon's concept of *bounded rationality* [174] may serve to explain one facet of the complex difference between the economic human and humanity as we understand it. According to Simon, the convoluted environment that humans inhabit limits the extent to which rationality can be applied since information about the environment may be incomplete due to factors such as *probability*. Simon posits that human rationality is therefore *bounded* and that some form of additional *emotional* input is required in order to enable an answer to be provided in such a variegated environment. Russell and Norvig's definition of a rational agent in [161] further highlights the potential breadth of information required in order for a rational agent to act optimally:

> "for each possible percept sequence, a rational agent should select an action that is expected to maximize its performance measure, given the evidence provided by the percept sequence and whatever built-in knowledge the agent has."

In the above quote, the term "*whatever built-in knowledge the agent has*" is exceptionally important and has far-reaching implications. Indeed, with regards to the decision-making of *homo sapiens*, it would seem that there is some other input to decision-making other than pure self-interested rationality. Research from psychology, behavioural economics and biology regarding this subject makes frequent use of game theory in order to determine if people act as rationally as the *homo economicus* agency model predicts and asks: if not, why not? The results of such research are often surprising with works such as [86], [57] and [159] illustrating that humans do not indeed always act rationally. In fact such research demonstrates that *homo sapiens* can act in complete opposition to the *homo economicus* model. For example: in certain situations, human beings will incur a financial loss to themselves in order to either establish or maintain cooperation with others instead of maximising individual utility [58].

Cooperation in MAS is considered as being integral to enable successful interaction between agents. Wooldridge highlights the importance of cooperation in [212], where it is stated that:

> "In order to successfully interact, these agents will thus require the ability to *cooperate*, *coordinate* and *negotiate* with each other, in much the same way that we cooperate, coordinate, and negotiate with other people in our daily lives."

Therefore, to develop truly autonomous systems a sensible course of action would be to first understand the mechanisms that establish and maintain cooperation amongst

individuals in human society and then to model such mechanisms so that they may be of functional use to autonomous agents secondly. The most notable work regarding motivations of cooperative behaviour in human beings has to be attributed to Fehr and Gächter who have published numerous papers on the subject advocating the role of *emotions* in enabling reciprocity of cooperation [57], [56]. In addition to the aforementioned authors there are other researchers who advocate emotion as being an important driver of cooperative behaviour. For example: in [164], Sanfey et al. show how unfair proposals in games from experimental economics cause increased activity in the anterior insula (associated with emotional responses) and dorsolateral pre-frontal cortex regions (associated with cognition) of the human brain. Furthermore, individuals with increased levels of activity in the anterior insula rejected so-called "unfair" offers more frequently. Similar results are observed by Quervain et al. in [40] but, unlike Sanfey et al., they specifically propose that *satisfaction* and even the anticipation of satisfaction, is the motivation behind some behaviours exhibited by players in such games.

Further psychological research into the role of emotion in decision-making has illuminated some interesting findings. Perhaps the most famous work regarding the role of human emotion in decision-making is that of Damasio who presents the *somatic marker hypothesis* in [35]. The hypothesis supports Simon's conclusions from his work on bounded rationality [174] i.e. Damasio proposes that cognitive processes regarding decision-making may become overloaded due to the complexity of the current situation being considered. In such circumstances *somatic markers*[1] are used to bias the cognitive procedures associated with decision-making by creating a *somatic state*. Damasio proposes that these somatic states help to identify possible, plausible options thereby simplifying the decision-making process. A number of works including [14], [15] and [13] test the decision-making of patients who have suffered bilateral damage to the ventromedial pre-frontal region of the brain; the area where it is believed that somatic markers are located. Such research concludes that damage to this area results in negative modifications to an individual's social behaviour and ability to decide advantageously on matters regarding their own lives. Remaining ignorant of the effects of emotion upon decision-making would therefore appear to be equally disadvantageous.

As stated, damage to the ventromedial pre-frontal region causes disruptions to an individual's social behaviour. It has been posited that emotions are an important determinant of such behaviour with works such as [70] arguing that emotional intelligence has a direct influence on the effectiveness of leadership. Other research including [200] concludes that the relationship between infants and adults is mutually regulated by emotional expression from both the care-giver and the care-receiver. It is therefore proposed that a major determinant of a child's development is related to the successful operation of this communication system.

Thus, if cooperative interactions in MAS are to be established and maintained, especially when the details of such interactions are not certain, then it is not just rationality

[1] Stimuli originating from the body requiring little to no mental cognition to provoke a reaction when experienced e.g. "gut" feelings of another upon first meeting.

and self-interest that can be considered. There is no need to limit ourselves in this way; pro-social, cooperative behaviour and the mechanisms that enable such behaviour can also be developed. This may yield interesting and useful insights, and perhaps increase the robustness of agent decision making. Therefore, in this thesis, I will attempt to computationally mimic emotional mechanisms since they clearly do have some form of impact on pro-social behavioural choices.

1.1 Research Question

Since emotions can affect people in a variety of ways, I need to select a particular focus for investigation. Consequently, I will attempt to augment the autonomy of agents so that cooperative behaviour may be established when rational, self-interested behaviour will lead to undesirable individual and total system results as in the *Prisoner's Dilemma* game [151], the *Dictator* game [98], the *Ultimatum* game [80] and the *Tragedy of the Commons* [83] (all of which will be discussed further in chapter 2). To achieve this, I consider the relatively unexplored role that emotions play in decision-making from an agent-systems perspective; specifically, what input do emotions have upon an agent's decision to either foster cooperation or destroy it? The salient research question can be therefore stated as:

> *"How should emotions be modelled functionally in a computational context so as to understand their impact upon the social interactions of agents engaged in public goods games? Which emotions should be modelled, what effects are observed when agents drive their interactions using these emotions and what is the impact on individual and system performance?"*

In answering this question I formulate a computational framework for modelling emotions in agents and draw upon various disciplines including computer science, social science, psychology, biology, philosophy, economics and mathematics. I also construct a novel MAS test-bed to facilitate the functional modelling of emotions using the framework developed and to allow for the acquisition of relevant data that is used to answer the research question posed. The framework for modelling emotion and the novel MAS test-bed mentioned will be described in detail in forthcoming chapters and the implications of the results obtained are analysed and discussed in detail. This consideration of the results obtained allows me to offer answers to the latter part of the question: "*what are the effects of the emotions modelled on the decisions made by agents and on a society of agents as a whole?*".

1.2 Overview of Topic

The study of emotion in computer science, particularly agent systems, has gained considerable attention of late as researchers continue to grapple with the inherent flaws

of the *homo economicus* model of agency. The aim of this thesis may appear to be the destruction of *homo economicus*, eradicating its influence not just from computer science but also economics yet I have no such intention. I believe that *homo sapiens* contains elements of the *homo economicus* model but postulating that the two models are one in the same with no other input from other models of agency is, in my opinion, a gross oversimplification. My intention in this thesis is to solely consider a model of the *emotional man* or *homo sensus* so that it may in future be married together with *homo economicus* to create a truer model of *homo sapiens*. The justifications for pursuing this aim are given in chapter 2 where the difference between the selfish predictions of *homo economicus* and the observed benevolence of *homo sapiens* in economic contexts is presented and discussed. Effectively, I am attempting to functionally model the mechanisms that enable agents to reach decisions when their rationality may be bounded as Simon posits with the concept of *bounded rationality* in [174]. By functionally modelling emotions in an explicit sense I hope to analyse their effects computationally so that I may experimentally verify what effects emotion has upon social interactions between agents when complete information about these interactions does not exist.

The question of what emotions should be modelled and how can they be modelled computationally is more complex than it first appears and a definition of emotion must first be presented. The concept of emotion is hotly contested and no real consensual definition has ever been formally agreed upon. Weschler [209], and Kleinginna and Kleinginna[105], demonstrate the difficulty in even reaching a consensual definition of the word itself, as a variety of terms exist for the multiple facets of emotion-related terminology. There is even a lack of consensus upon what counts as an emotion with some classifying feelings such as *hunger* as emotions whilst others do not. With respect to what an emotion is, James in [94] views emotions as being purely physiological i.e. "fear" is a lexical token assigned to describe a collection of bodily reactions such as quickened heart rate, profuse sweating, weakness of limbs etc. produced in response to external stimuli. Kenny expands upon this view in [103] and posits that an emotion is composed of three attributes namely a stimulus, an intentional response and a physiological symptom. He states that in any case, one of these attributes may be missing but at least two are needed e.g. fear may be triggered by the perception of a spider and may then cause sweating and/or rapid movement away from its vicinity. Other psychologists take the view that emotions facilitate evaluations of behaviour that are expressed when a particular situation arises. Such an evaluation has the potential to modify the behaviour of the same individual so that a similar or different action is performed if that situation occurs again [10]. Like Keltner and Gross [102] and Frijda in [67], I adopt a functional view of emotion and have therefore used such an approach to model them. The emotions modelled in this thesis have been selected to illuminate particular aspects of cooperative behaviour and will be discussed later in chapters 6, 7 and 8. The emotions are: anger and gratitude (chapter 6), admiration (chapter 7) and hope (chapter 8). I do not consider physiological aspects of emotions in this thesis since they are not generally applicable to

MAS. The question of whether emotions can be modelled satisfactorily in the absence of physiological considerations is discussed in detail in chapter 3 along with other pertinent philosophical questions.

It is important to draw attention to the fact that there exists no generally accepted framework for modelling emotion, especially from a computational standpoint. In fact, it would appear that every computer scientist who attempts to use a computational model of emotion invents their own for their own purposes; there is no standard. One of the other major aims of this thesis is to develop a general framework for modelling emotions computationally in the context of agent systems by using solid psychological models of emotion as a basis. I therefore intend to establish this framework as the common framework upon which others may model emotions in a functional manner in a computational context. Whilst there have been numerous attempts to formalise psychological models of emotion in agent specific languages; arguably the most successful of these attempts has been made by Stuenebrink et al. in papers such as [186] and [187]. The authors model emotions as *fluents* i.e. their associated intensities are capable of increasing and decreasing enabling emotions to become activated and deactivated autonomously and dynamically. The most salient contribution of this work is its attempt to prescribe heuristic functions with respect to an agent's decision-making when emotions are activated. Steunebrink et al. only cover a handful of emotions however, so there is much research to be conducted regarding the remaining emotions. Steunebrink's framework and other notable examples are considered in chapter 4 and salient aspects of such logical formalisms and computational implementations are identified. By doing this I aim to both develop the computational framework used to functionally model emotion that is proposed in this thesis and illustrate how this framework relates to those proposed in the previous work of others discussed in chapter 4.

1.3 Contribution to Knowledge

As mentioned in section 1.1 the aim of this thesis is to develop and functionally model emotionally-inspired mechanisms of decision-making for agents. Through this I hope to provide a greater understanding of how emotions can impact upon decision-making and social actions between agents in particular public goods games. More precisely, the aim of this thesis is to bridge the gap between *homo economicus* and *homo sensus* so that a model of agency that mimics *homo sapiens* much more closely can be developed and refined. This will be achieved by using experimental evidence from other disciplines to identify the emotions elicited in particular public goods games and what their effects are upon societal interactions in this context. The emotions identified and their associated attributes will then be modelled computationally using the research mentioned earlier and framework I have developed. These emotions will then be implemented in agents that comprise a MAS to determine what effects the emotions modelled have upon social

interactions and why they have these effects. This question comprises a number of sub-questions due to the fact that different emotions are modelled and different effects are observed. As such, there are three chapters: chapter 6, chapter 7 and chapter 8, devoted to providing answers to this question. Each of these chapters discusses the modelling of a single or related pair of emotions using the framework developed and the effects of these emotions upon the societal interactions of agents.

Ultimately, this thesis contributes the following knowledge to the current state of the art:

- An experimentally verified method that enables MAS designers to model emotions functionally in a computational context.
- A demonstration that emotions can (and arguably should) be an integral part of an agent's decision-making mechanism.
- A demonstration that emotionally-inspired decision-making mechanisms can be beneficial to an agent system.
- An understanding of the effects of the emotions modelled on both a micro and macro scale given differing initial system conditions.
- An understanding of the effects on both a micro and macro level when combinations of more than one emotionally-inspired decision-making mechanism is considered within a MAS.

1.4 Thesis Structure

The remainder of this thesis is organised into eight chapters, descriptions of which can be found below. Due to the interdisciplinary nature of the work undertaken a considerable literature review has been produced which spans chapters 2, 3 and 4. The original work comprising chapters 5, 6, 7 and 8 then draws upon various aspects of this literature review.

Chapter 2 provides a review of literature comparing the validity of game theory predictions of public goods game outcomes and the outcomes of such games when played by humans. From here, I review literature that discusses what it is about human players that may drive this deviation from rational predictions namely, emotion.

Chapter 3 is particularly concerned with whether emotions are capable of driving intentional behaviour and whether there is a dichotomy between "emotional" and "rational" decision-making. Answers to both these questions are provided in this chapter following a review of relevant philosophical and psychological literature.

Chapter 4 presents research regarding major psychological models of emotion, logical

formalisms of some of the aforementioned psychological models of emotion and existing computational implementations of emotion. By taking such work into account I aim to both contextualise the work presented in this thesis and also to identify salient aspects to be carried forward when developing the emotion modelling framework presented. Furthermore, the intention of this chapter is to distinguish an appropriate method for testing how the emotions modelled and implemented using the proposed framework affect societal interactions in a computational manner.

Chapter 5 outlines the research agenda of this thesis in detail along with specifics of the test-bed and the computational framework for modelling emotion used in the remaining chapters. Novel terminology that pertains to the emotional modelling framework and the thesis in general is also presented and defined in this chapter.

Chapter 6 details the simulations constructed and run to experimentally investigate the functional modelling of *anger* and *gratitude*. These simulations also serve to ascertain the effects of these emotions upon societal interactions when emotional agents are pitted against notable strategies from Axelrod's *Evolution of Cooperation* tournament [7]. The results obtained by these simulations first facilitate a discussion of the importance of anger and gratitude in the context of public goods games. Secondly, the results are used to comment upon whether altering how easy it is to activate these emotions enables an agent to be more "successful" than the "tit-for-tat" (TFT) strategy of Axelrod's tournament fame. Finally, the chapter discusses how changing the rate at which these emotions are activated affects the agent's ability to establish and maintain cooperative behaviour and produce fair systems. The work constituting this chapter was published in the 4th International Conference on Agents and Artificial Intelligence (ICCART 2012) [112].

Chapter 7 augments agents implemented with anger and gratitude with *admiration*. This emotion allows agents to copy the rate at which other individually successful agents activate anger and gratitude. By doing this, it is possible to determine whether there is one configuration of activation rates that becomes more prevalent than others given different initial simulation conditions or how quickly admiration is elicited in others. Essentially, this chapter investigates whether one social norm with regards to the rates of anger and gratitude emerges from the system or whether different social norms emerge given different initial circumstances and why. The work in this chapter was published in the 13th International Workshop on Multi-Agent Based Simulation (MABS 2012) [113].

Chapter 8 introduces *hope*, whose effect is to destabilise CC between opponents in public goods games when emotional social norms already exist. Again, different rates at which anger and gratitude are activated are scrutinised with respect to their effects upon social interactions in conjunction with how these activation rates interact with hope.

Chapter 9 provides a complete summary of the work performed and presents the key contributions that this thesis has made. Further research that may be carried out using the work presented is also suggested in this chapter along with a discussion of current world events relevant to this work.

Chapter 2

Emotions and Human Behaviour

In this chapter I give an extensive review of background literature that serves to justify why this thesis attempts to both functionally model emotion in agent systems and determine the effects of emotion upon societal interactions in context of public goods games. To achieve this, two general points are made: sections 2.1, 2.2 and 2.3 provide a discussion of literature regarding the predictions of game theory when applied to public goods games and how these predictions do not apply when such games are played by human beings. The issues considered in these chapters have a significant impact upon the design of experiments that will be constructed and run in later chapters of the thesis. Sections 2.4 and 2.5 then provide a review of literature that identifies in what particular ways humans do not adhere to the predictions of game theory and suggests why the behaviour of humans opposes the "rational" strategies suggested. The results of these works both serve to justify and to guide the design of the emotional model proposed.

To demonstrate how the behaviour of *homo sapiens* and *homo economicus* does not equate in the context of public goods games, section 2.1 first presents the salient social dilemmas that follow from adopting the rationale prescribed by the *homo economicus* model of agency used in game theory. General anthropological and experimental economic evidence that deviates from the predictions made by the *homo economicus* model of agency is then considered in section 2.2. The intention of this section is to make clear the disparity between *homo economicus* and *homo sapiens* in a general public goods game context. The disparity observed centres around the fact that, instead of maximising rational self-interest, humans in most societies are willing to help others even when some form of loss (in time, food, money etc.) is incurred. Given that this discussion refers to the behaviour of humans in general public goods game situations I then lead into a focused discussion of literature regarding the behaviour of human beings in the Prisoner's Dilemma game in particular (see section 2.3). This section again illustrates how rational self-interest is not the primary motivation behind the intentional behaviour exhibited when human players are engaged in such games. Rather, there would appear to be something else other than monetary income that players attempt to maximise and it is this "something" which fosters cooperation in situations where rationality would prescribe defection.

At this point it is important to clarify that, in this thesis, I propose that behaviour may be divided into two types: *unintentional* or *involuntary* behaviour and *intentional* or *voluntary* behaviour. I define unintentional behaviour as being actions that are performed by an agent which the agent has no control over (in the context of emotion, such behaviour is usually physiological in nature). For example: sweating when fearful, pupil dilation when happy, clenched jaw when angry etc. I define intentional behaviour as behaviour demonstrated by an agent that the agent has deliberated about and intended to perform prior to its execution to bring about some desire. For example: punishing another in response to anger, rewarding another in response to gratitude etc.

Section 2.4 then proposes an explanation for the cause of the dissonance between *homo economicus* and *homo sapiens*. The literature introduced in this section discusses what the additional factor mentioned in section 2.3 that humans attempt to maximise could be. By considering various strategies that are capable of establishing and maintaining cooperation in public goods games like the Prisoner's Dilemma, I posit that *emotion* motivates and drives the usage and adherence to the strategies discussed. Section 2.5 then expands upon section 2.4 by reviewing literature that outlines the explanatory power of emotion with respect to the establishment and maintenance of intentional, cooperative behaviour amongst human beings.

2.1 Game Theory and Social Dilemmas

Game theory is a mathematical theory concerned with the study of interactions between self-interested or rational agents [20]. According to Wooldridge [212], game theory is used by economists to study and understand the interactions amongst economic entities in social reality. This section examines some predominant games commonly used in game theory that are of interest to this thesis. The common theme between these games is that there is some concept of social resource that requires consideration when planning strategies. With reference to these strategies, the most rational strategies for such games either result in dilemmas or unfavourable pay-offs for at least one player that can be observed in many human societies today.

2.1.1 The Dictator and Ultimatum Games

The Dictator and Ultimatum games do not create an explicit social dilemma quite like the Prisoner's Dilemma and the Tragedy of the Commons (described in sections 2.1.2 and 2.1.3 respectively). However, a discussion of this game is included as some understanding needs to be had of the rational actions prescribed by the *homo economicus* model in the game so that they may be juxtaposed with the actions of actual human players later in the chapter. It should be noted that the Dictator game in particular presents a very clear example of rational behaviour due to the conditions of the game itself.

The Dictator game was originally developed and used by Kahneman et al. in [98] and is not strictly a "game" as defined by game theory since, even though two players

are required to play, the role of one is entirely passive. The game is outlined as such: one player, the proposer, decides how to split a given resource (usually some monetary amount) with the other player, the responder, and the responder has no choice but to accept. The *homo economicus* model of agency would dictate that the most rational course of action would be for the proposer to not offer anything thus securing the entire resource for him/herself whilst the responder receives nothing. In practice however, experimental evidence shows that very few people adhere to such a strategy even across different cultures and when the monetary amount available to split is higher as Engel [54], Cooper and Dutcher [32], Slonim and Roth [176] and Cameron [27] report.

The Ultimatum game permits the responder to have a more active role and was created and utilised by Güth et al. in [80]. The game initially proceeds in the same way as the Dictator game: the proposer decides how some resource should be split. The difference between the Ultimatum and the Dictator game is that the proposal may or may not be accepted by the responder. If the responder does not accept then both players receive nothing, otherwise the amount proposed is deducted from the total stake and given to the responder and the proposer receives the remainder. Therefore, the most rational course of action for a proposer is to offer the minimum amount permissible. For the responder, the most rational course of action is to accept any proposal offered since gaining something is better than gaining nothing. Although, as with the Dictator game, when people play this game the rational strategy is rarely employed (see [81], [158] and [183] for experimental evidence that supports this claim). This would suggest that economic gain is not the sole concern of participants.

2.1.2 The Prisoner's Dilemma

The Prisoner's Dilemma game is familiar to most people, even if they are not explicitly aware of its name, details or predictions. The game was first devised by Flood and Dresher whilst working for the RAND (Research and Development) corporation and was formalised by Tucker in 1950 with the Prisoner's Dilemma name and prison sentence pay-offs [151]. The general rules of the game are simple: two players (who are not allowed contact with each other prior to playing) have the choice of two actions: cooperate or defect. In the formalised prison sentence version of the game, the players are suspected of a crime that they committed together. Each player is questioned in separation to each other and is asked if the other prisoner was involved in the crime. To stay quiet and not confess anything is termed as a "cooperation" (as the prisoner cooperates with his accomplice) whilst confessing that the other prisoner was involved in the crime is termed as a "defection" (as the prisoner defects against his accomplice).

Play occurs simultaneously over one round so neither player has a chance to react and reward/punish their opponent. There are four possible outcomes to the game as shown in table 2.1 and the ordering of these outcomes creates the dilemma of the game's namesake. What is important about the game itself is that it is the joint action of the players that determines the final outcome for both. So, whilst an agent can act in its

TABLE 2.1: Prisoner's Dilemma pay-off matrix.

		Player i	
		Co-op	*Defect*
Player j	*Co-op*	3_i, 3_j	5_i, 0_j
	Defect	0_i, 5_j	1_i, 1_j

own self-interest, the consequence of the individual's actions upon the total score of the system may also be considered. Note that in table 2.1, the values represent the utility earned by an agent if a particular joint outcome occurs i.e. 5 is the best outcome for an agent whilst 0 is the worst.

From the individual point of view of player i, the preference ordering of outcomes according to a rational model of agency is: $\mathrm{D}_i\mathrm{C}_j > \mathrm{C}_i\mathrm{C}_j > \mathrm{D}_i\mathrm{D}_j > \mathrm{C}_i\mathrm{D}_j$. Therefore, the best outcome for player i is to defect whilst player j cooperates. However, if player j is also rational then j would also defect causing both players to achieve their third most preferred outcome. As can be seen, cooperation carries with it a risk of achieving the worst possible outcome. So in order for i to both maximise its individual score and safeguard against being exploited, it should always defect, no matter what j does.

If on the other hand the total system score of each outcome is considered then, from the perspective of player i, the preference orderings may be rearranged: $\mathrm{C}_i\mathrm{C}_j > \mathrm{D}_i\mathrm{C}_j > \mathrm{C}_i\mathrm{D}_j > \mathrm{D}_i\mathrm{D}_j$. In this context, CC is the most preferred outcome as this yields a total of 6 whilst DD is the worst outcome as a system total of 2 is achieved. The dilemma here is that if both players are rational, the best action to employ is to defect but as explained, this results in a sub-optimal outcome for both players both in an individual and a total system context. CC on the other hand produces a much better outcome for both the system and the players involved but, if you intend to cooperate, how do you ensure that your opponent will do the same and so ensure that you do not receive the worst possible outcome? This is the true dilemma that the players face.

There also exists an iterated version of the Prisoner's Dilemma where it is possible for a player to reward/punish an opponent in a subsequent round for cooperation/defection in a previous round. If, however, the number of iterations is finite and both players are rational then each will expect the other to defect on the final round n as there can be no retaliation in subsequent rounds. Therefore, the last "real" round is *n-1* and again if both players are rational then both players would be expected to defect on this round as both will have defected in round n regardless. This argument continues backwards until the initial round meaning that both agents will mutually defect for the entire game regardless of the presence of subsequent rounds. Such reasoning is termed as the "backwards induction paradox" and is outlined in [144]. The existence of this paradox is the primary argument for why DD occurs in the iterated version of the game. As stated however, the number of rounds must be known by the players in order for the

effect to be manifest. Consequently, the iterated version of the game is normally played over a number of rounds not known to the players.

2.1.3 The Tragedy of the Commons

The social dilemma presented by the Prisoner's Dilemma in section 2.1.2 is not just confined to that game. In the Tragedy of the Commons [83], a scenario is posed where a number of farmers each share a finite plot of land. Upon this finite plot, farmers may graze cows and these cows in return provide some kind of utility to the farmer. For example: cows are able to produce milk or be butchered for meat etc. The produce of the cows can then be sold, earning the farmer money. Now, supposing each farmer has one cow and there is an abundance of land available on the plot, if a farmer is rational, he may introduce another cow onto the land so that it may also graze, thus doubling the farmer's utility. The problem though, is that this removes space for other cows on the plot; if this behaviour is extrapolated then the social resource i.e. the finite plot, will eventually become so overgrazed by the cows that it will become unusable. If this should occur, the individual actions of each farmer will be of extreme detriment to both their individual long-term interests and the society of farmers as a whole. Again, rationality creates an outcome that is both of detriment to the individuals in a society and to society as a whole.

2.2 *Homo Sapiens* and Irrationality

Following on from the discussion of how the *homo economicus* model of agency creates the social dilemmas outlined in section 2.1, this section discusses literature that provides experimental evidence refuting the notion that courses of action posited by this model are followed in practice. The literature presented in this section proposes that the *homo economicus* model is neither sufficient nor correct in the majority of cases for predicting the behaviour of *homo sapiens* in public goods games. Instead, the most likely behaviour of *homo sapiens* in this game context is to cooperate with others rather than to defect.

2.2.1 Anthropological Observations

First, cooperation in human societies from an anthropological standpoint and independent of the game contexts discussed in section 2.1 is considered. Research by Kaplan et al. [101] theorises that cooperation between genetically unrelated human individuals is crucial to mechanisms that ensure human longevity. This hypothesis is based upon the observation that *homo sapiens* exploit high-quality consumables that are difficult to acquire and, therefore, a high degree of knowledge, skill and coordination is required. To attain such abilities an individual must devote a large amount of time to education since the longer a human being can spend learning, the greater the pay-off for both the individual and his/her community when this knowledge is utilised. Due to this, a selection pressure is expected upon genes that promote extended juvenile states since

learning may be fostered considerably more during this time; herein lies the crux of the argument. Whilst in a juvenile state, human beings are dependent upon others for help with regards to food, knowledge etc. so cooperation from others is required. Consequently, a selection pressure is also expected upon genes that promote cooperation since, the more cooperative the population, the longer juveniles can spend in education honing their skills and the more juveniles that can be immersed in this education. This ensures a greater pay-off for both the population as a whole and for the individuals that comprise it when juveniles who have emerged from extended periods of education make use of their highly trained skills.

The propensity for cooperation in human beings is also outlined by Hill in [89] where a study of the hunter-gather Ache tribe from eastern Paraguay shows that significant rates of cooperation exist during all phases of food foraging. Hill posits that the Ache share all foods acquired because of the ubiquitous cooperation that exists at all foraging phases. Hill also notes that the three resource types that account for the most foraging time (honey, fibre and meat) are also the resources that are the most widely shared between foraging bands. Therefore, there would appear to be a positive correlation between time devoted to cooperation and amount of resource shared. A further important point made in [89] is that cooperation between individuals also exists for tasks that do not involve food acquisition. For example: Hill observed that Ache women also provided child-care for other families amongst other social activities.

Note that the examples of cooperation reported in these pieces of research do not make it clear whether cooperation is or is not rationally advantageous from the perspective of the individual in the long term. It may be that one helps another since the cooperator knows or hopes that this cooperation will result in a greater pay-off to themself at some later date. Such *reciprocal* strategies are discussed further in section 2.4. It is important to remember that cooperation and selfishness are not mutually exclusive concepts.

2.2.2 Observations in Experimental Economics

Numerous economic experiments involving human beings also demonstrate the rejection of predictions made by the *homo economicus* model of agency. Henrich et al. [86] presents extensive evidence derived from 15 different non-industrialised societies playing Ultimatum games. The evidence provided categorically refutes the idea that *homo sapiens* act in accordance with the *homo economicus* agency model since the lowest mean stake shares exceeded 25% of the total stake whilst the highest mean stake shares equalled 58% of the total stake.

Rates of cooperation in industrialised societies were investigated by Roth et al. [159] who devised two games to investigate the cooperation phenomenon exhibited by *homo sapiens*: the first is a standard Ultimatum game, whilst the second is a modification of the standard Ultimatum game entitled the *market* game. In the market game, nine buyers make independent offers to a seller for an object and the seller has the opportunity

to accept or reject the highest offer made (if more than one buyer offers the highest price then a winner is selected by lottery). If the seller accepts then the seller earns the highest amount and the successful buyer earns the difference between the object's value and the price they paid i.e. if I bid \$995 for an object that is worth \$1000 and I submitted the highest bid then I would earn \$5; all other buyers receive \$0. If the seller rejects the highest offer then all buyers receive \$0. In the experiments, participants are only able to make offers in multiples of 5.

With respect to the market game, optimal strategies were found by all societies tested no earlier than round 3 and no later than round 7. When playing this game it appears that players act in accordance with game theory predictions but no discussion is given to why this may be. A potential answer may be that learned behaviour in the context of a game overrules habitual behaviour in social reality. Essentially, people may learn the game and strategies that will allow them to win; the principles behind such strategies may differ from the principles of how to act in similar, "real" situations. However, with respect to the Ultimatum game the results are markedly different since the optimal strategies of offering 0 or the next smallest amount constitute less than 1% of the data in any society examined. In all societies tested, offers were closer to half the amount to be split and low offers are rejected at substantially higher rates than high offers in all societies. The explanation offered for this result is that of *fairness* i.e. each society tested has some social norm regarding what is a fair offer and those offers that do not come close to matching this are rejected. Roth et al. note that the results observed appear to be consistent with results from other Ultimatum bargaining games conducted by Güth et al. [80], Prasnikar and Roth [152] and Roth [158]. The salient explanation provided by Güth et al. and verified by the results obtained by Roth et al. for the rejection of offers by human responders is summed up by the proposed reasoning process of both players:

> "The typical consideration of a player 2 (responder) in an easy game seems to be as follows: 'If player 1 (the proposer) left a fair amount to me, I will accept. If not, and I do not sacrifice too much, I will punish him by choosing conflict'. Correspondingly, a player 1 typically will argue like: 'I have to have at least an amount $c - a_1$ for player 2 so that he will consider the costs of choosing conflict too high'. "

In [73], Gintis provides yet more evidence that human beings do not act in accordance with rational models of agency. This literature review provides details of two types of experiments under which the congruence of the *homo economicus* model to reality is tested: individual choice behaviour and strategic interaction. Having already considered strategic interactions, I will focus upon Gintis' review of literature regarding individual choice behaviour.

A number of factors are examined with respect to individual choice behaviour, the first of which are time inconsistency and hyperbolic discounting. After a consideration

of literature regarding these issues, Gintis proposes that people do not act in accordance with economic theory with respect to time consistency i.e. he claims that people tend to favour short-term gains even when they entail long-term losses. Research by Lowenstein and Sicherman [116], [119] also provides support for the argument that agents have different rates of discounts for different rates of outcomes. The example situation given in [116] is that of a person who is offered two free dinners for themselves and a friend. One meal is to be had next weekend and the other is to be had in two months time. The recipient also has a choice of two restaurants to eat in: a fast food outlet and the fanciest restaurant in town; one meal is to be eaten in each. Rational economic theory would dictate, assuming that discount rates are positive, that the fancy meal should be had first and the fast-food meal second however, people who were offered such a deal made the opposite choice as they preferred the anticipation of eating in the fancy restaurant. Similar results are obtained by Lowenstein and Sicherman in [119].

The second factor investigated with respect to individual choice behaviour is judgement under uncertainty; Kahneman et al. supplies the leading piece of literature concerned with this topic [100] (this work has already been mentioned in chapter 1). Experimental results from this work show that rational considerations such as statistical probabilities are not considered when people are asked to determine the career of someone based upon a brief description. For example: when people were asked to determine if someone who has a good sense of humour and likes to entertain friends and family is either a professional comedian or a clerical worker, people are more inclined to say that they are a professional comic. Statistically speaking, such a conclusion is bogus since there are far more clerical workers in the world than there are professional comics therefore, it is much more likely that the person described is a clerical worker.

Finally, Gintis discusses loss aversion and status quo bias in the context of individual choice behaviour. Experimental evidence from Helson [85] and Tversky and Kahneman [204] concludes that people are about twice as adverse to taking losses as they are to enjoying an equal level of gains. Implications of loss aversion are the endowment effect which has been principally investigated in [99], whereby people demand much more to give up an object than they would to acquire it (as defined by Thaler [197]). A further implication observed is that of the framing effect investigated by Tversky and Kahneman [202]. This is where, given the choice of two scenarios whose results are the same, a different description of the scenarios causes people to alter their choice. This is shown in [202], where two sets of participants are able to choose between two alternatives to solve a problem. The wording of the solutions is altered depending upon the set of participants but the calculations and results of the solutions are identical. 72% of the first set of participants chose the rational first alternative whilst 28% chose the other. 22% of the second set of participants however chose the rational first alternative whilst 78% chose the other. If human beings were indeed completely rational then both groups of participants would have yielded similar results (the majority of participants in both sets would have picked the rational first solution), that they did not do so demonstrates

the framing effect. Framing effects are found regularly in experiments of this kind.

The last implication of loss aversion is the status quo effect again investigated by Kahneman [99]. Samuelson and Zeckhauser originally defined status quo bias as being a preference for the current state of affairs as opposed to any other [163]. To clarify this, Kahneman et al. give an example of a friend who bought a bottle of wine originally worth \$10 and which could now be worth up to \$200 at auction. This friend now neither wants to sell the wine (earning him a considerable profit) or buy the wine at the new price therefore, he is happy with the current status quo. Samuelson and Zeckhauser conducted a range of experiments in order to test this concept and found that:

- Given a current situation and the opportunity to change to other alternatives, participants will generally stick with their current situation even if an alternative would endow the participant with increased utility.
- If one of the alternatives is designated as the status quo rather than the participant's current situation the alternative became more popular.
- Participants perceived the advantages conferred by the status quo to increase as the number of alternatives increased.

The topic of irrational human behaviour in the context of public dilemma games is notably tackled by Dawes and Thaler in [38] where a comprehensive discussion of the topic of cooperation in such games and related works is provided. With particular reference to the Dictator game, Engel provides an extensive meta-analysis of 616 treatments in [54] and reaffirms the results presented in this section along with an elegant presentation of results illustrating how variations in players such as age, the generosity of players etc. affect the offers of proposers. The number of result sets that have been collected and analysed from within the context of one particular public goods game are staggering, so attempting to comment upon all result sets for all public goods games would be a massive undertaking. Taking this into account, it would seem that the most sensible course of action would be to concentrate on one public goods game in particular and attempt to both understand and explain how and why cooperation emerges between human players.

2.3 Cooperation within the Prisoner's Dilemma

Previously in section 2.2, I endeavoured to illustrate the large amount of research concerning human "irrationality" in the context of a number of public goods games. The culmination of this consideration was a conclusion that it would be best to focus upon one of these games so that a rigorous analysis of why human players act "irrationally" can be performed. Consequently, in this section I shift my focus onto a specific consideration of the Prisoner's Dilemma game as the amount of literature regarding the study of cooperation in context of this game is extensive. Also, the Prisoner's Dilemma allows

me to ascertain the extent to which players value their own self-interest over the good of the system, as self-interest carries with it a serious risk of producing unfavourable outcomes for both players. The aim of this section is to discuss experimental evidence derived from Prisoner's Dilemma games that augments my previous assertion that the *homo economicus* model of agency is insufficient to describe and predict the behaviour of *homo sapiens* both in and out of market environments.

One of the most interesting studies regarding the behaviour of human beings when playing the Prisoner's Dilemma game has been undertaken by Maxwell and Ames in [125] where a strong and a weak free-rider hypothesis are experimentally tested. Anonymous human players are given a resource and told that they may either invest privately for a fixed return or contribute to a public good whose payout is dependent upon investment of others. "Free-riding" is defined as an absence of contribution to group investment, therefore strong free-riding occurs when voluntarily contribution is absent and weak free-riding occurs when a sub-optimal number of individuals contribute. The results show that the strong free-rider hypothesis was consistently rejected, instead the weak free-rider hypothesis was supported: levels of cooperation never reached 100% but the percentage of cooperation never dipped below 28% and peaked at 84%.

Independent work conducted by Frank et al. [64] attempts to investigate what effect social interaction has upon cooperation and defection rates in the Prisoner's Dilemma game. In this investigation, three participants and three types of experiment were used. Details of the experiment types are described below:

- Unlimited - Participants were allowed to converse for 30 minutes prior to the game starting and were allowed to make promises to cooperate, although the enforced anonymity of participants rendered such promises unenforceable.
- Intermediate - Participants were allowed to converse for 30 minutes prior to the game starting and were not allowed to make promises regarding cooperation.
- Limited - Participants were allowed to converse for 10 minutes prior to the game starting and were not allowed to make promises regarding cooperation.

After conversation, participants were taken to separate rooms and told to write down their decision to cooperate or defect on a piece of paper. The effect of social interaction was shown to increase the amount of cooperation observed with unlimited games providing the highest rates of cooperation, intermediate games provoking less cooperation than unlimited games but more cooperation than obtained from limited games. Therefore, there is experimental evidence that indicates that social interaction and commitment devices seem to play a part in enabling and ensuring cooperation between human beings.

Similar results are again observed by Frank et al. in [65] where groups of three players are again allowed to converse with each other for 30 minutes before a game is played. As in [64], anything may be discussed within this pre-game interaction time, including

play strategies. The primary aim of the experiments was to determine if participants could accurately predict the behaviour of opponents after this brief period of interaction. Along with the decision to cooperate or defect, participants were also asked to predict the behaviour of both their opponents and assign a rating of confidence between 50 and 100 to their decision (50 = prediction was no better than chance, 100 = prediction is made with complete certainty). The results from these experiments again categorically refute the behavioural predictions of pure rational agency with 73.7% of participants cooperating and only 27.3% defecting. With respect to likelihood of cooperation, participants rated this likelihood higher (81.3) than it actually was (73.7). Of the 161 participants expected to cooperate, 130 (80.7%) did and of the 37 participants expected to defect, 21 (56.8%) did. Frank et al. calculated the expected average accuracy of predictions if calculated by chance alone arriving at a figure of 64.8% whereas the average accuracy of participants equalled 76.3% (the likelihood of arriving at this number by chance is less than 1 in 1000). Subjects were apparently more confident of their predictions regarding cooperation (average confidence rating = 85.7) than their predictions regarding defection (average confidence rating = 76.5). As before, social interaction seems to improve not only cooperation rates between participants but also predictions of behaviour allowing commitments to be maintained after their establishment.

Perhaps the most comprehensive body of work devoted to the topic of cooperation and how social interaction increases rates of cooperation in context of the Prisoner's Dilemma game has been undertaken by Sally in [162]. In this paper, a meta-analysis of 35 years of Prisoner's Dilemma experiments is presented that unequivocally demonstrate that human participants are not primarily motivated by rational self-interest. One of the most interesting results provided by Sally pertains to comparing strategies of participants that play for real money against those who did not. The *homo economicus* model would suggest that, when real money is being played for, the rate of defection would increase as the reward offered for defecting is tangible to the participants. A participant's awareness of playing a game may also cause defection rates to increase since, in a game, the objective is to win, not to lose or draw. Surprisingly, rates of defection decrease when participants play for real money and Sally's analysis also shows that conversation, a factor that should not affect the decision-making of a self-interested individual, increases cooperation rates by as much as 40% compared to games with no conversation. It is also shown that promises to cooperate mean more when they are agreed upon through conversation rather than text-only messages which appear to have no significance.

2.4 Why Cooperate?

In section 2.3 I discussed experimental evidence which demonstrates that the behaviour of human beings when playing the Prisoner's Dilemma is not in keeping with rational behaviour prescribed by the *homo economicus* model of agency. It has been proposed

in section 2.3 that the surprising prevalence of cooperation is due to the effects of conversation or rather, social interaction. Participants who interacted directly with other group members were more than likely to keep their promises of cooperation than those who did not.

Therefore, in this section I attempt to posit that the *emotional* component of social interaction motivates above-average rates of cooperation in social-dilemma games by considering a number of human-centric strategies. These strategies experimentally demonstrate how cooperation can be enabled by such strategies in Prisoner's Dilemma games. In the context of these human-centric strategies, I intend to discuss how emotion is used to ensure that these strategies are utilised and adhered to. By pursuing these lines of enquiry I hope to argue that emotion should be considered as an important factor in the enabling of cooperative strategies. Consequently, serious consideration should be given to the modelling of emotions in agent systems so that emotional models of agency that promote cooperation within the systems they are used in can be moved towards and produced.

2.4.1 Reciprocity

The effects of reciprocity have been studied most notably by Fehr and Gächter in [57] where they propose that reciprocity has powerful implications for economic domains and is an important determinant in the enforcement of social contracts and norms. They define two types of reciprocity in this work: positive reciprocity (discussed in section 2.4.1.2), which is concerned with cooperative reciprocal behaviour and negative reciprocity (discussed in section 2.4.1.1), which is concerned with retaliatory behaviour. The paper goes on to discuss how under certain conditions, specifically in competitive market environments where incomplete contracts may exist, reciprocity dominates as a strategy instead of self-interested, rational behaviour. A number of laboratory experiments are considered, the first of which is a public goods experiment where positive and negative reciprocity can exist. To maintain contributions to the public good, players must demonstrate reciprocal behaviour (positive reciprocity). It is more than likely, however, that some participants will free-ride as they will be motivated by self-interest. On observation that another player is free-riding, a player may then free-ride themselves causing a reduction in the free-rider's pay-off; this is the game's form of negative reciprocity.

An extension to this experiment is then proposed whereby participants are informed of how other participants are playing. Participants are then given the opportunity to punish others by placing a fine upon free-riders at a cost to the punisher (important since if it costs nothing to punish then self-interested free-riders would have no qualms in punishing cooperative individuals). Provision of punishment allows reciprocators to induce reciprocal behaviour in free-riders and its effect is dramatic with Fehr and Schmidt in [59] presenting evidence that a minority of reciprocal participants are capable of converting a majority of defectors into cooperators.

2.4.1.1 Negative Reciprocity

The work presented in [57] does not provide an answer as to what motivates reciprocal subjects to reciprocate. An answer regarding negative reciprocity is posited by Fehr and Gächter in two separate papers, the first of which, [56], presents evidence to support the hypothesis that *emotions* are guarantors of credible threats. In this work, Fehr and Gächter set up public goods games with four types of treatments with varying degrees of punishment: the more severe, the more costly it is to invoke:

- Partner treatment *with* punishment: 6 groups of 4 participants play 10 repeated rounds of the game and punishment of free-riders is possible.
- Partner treatment *without* punishment: 6 groups of 4 participants play 10 repeated rounds of the game and punishment of free-riders is not possible.
- Stranger treatment *with* punishment: 6 groups of 4 participants play 10 repeated rounds of the game and punishment of free-riders is possible. Participants are anonymous and, after each round, participants are randomly allocated to other groups and play resumes.
- Stranger treatment *without* punishment: 6 groups of 4 participants play 10 repeated rounds of the game and punishment of free-riders is not possible. Participants are anonymous and, after each round participants are randomly allocated to other groups and play resumes.

The results obtained from the investigation indicate that punishment has a positive effect upon rates of cooperation and a link between severity of punishment and degree of free-riding observed appears to exist. Fehr and Gächter attempt to determine what the proximate source of punishment is and propose that emotions play a key role in motivating punishment as other suggestions appear to fall short of explaining the results obtained. They cite Hirshleifer [90] and Frank [62] as being proponents of the idea that free-riding elicits strong negative emotions amongst cooperators and these emotions instigate a desire to punish free-riders. To ascertain the emotions experienced by participants when encountered with free-riders, Fehr and Gächter asked participants to indicate the intensity of negative feelings towards a free-rider or towards themselves if they free-ride. The results show that:

- More intense negative emotions are elicited when cooperators contribute more and free-riding occurs.
- Less intense negative emotions are elicited when cooperators contribute little and free-riding occurs.
- Participants expect more intense negative emotions directed towards themselves if they free-ride whilst others contribute large amounts.

- Participants expect less intense negative emotions directed towards themselves if they free-ride whilst others contribute small amounts.

Fehr and Gächter marry their hypothesis regarding emotion with the results obtained from the public goods experiments in three ways:

- "*...if negative emotions trigger punishment one would expect that the majority of punishment activities is executed by those who contribute more against those who contribute less. This is the case both in the Stranger- and the Partner-treatment. Between 60 and 70 percent of all punishment activities follow this pattern.*"
- "*...remember that non-strategic punishment increases with the size of the negative deviation from the average. This is exactly what one would expect if negative emotions are the cause of the punishment because negative emotions are the more intense the more the free rider deviates from the others' average contribution.*"
- "*if negative emotions cause punishment, the fact that most people are well aware that they trigger strong negative emotions [...] in case of free riding renders the punishment threat immediately credible. Therefore, we should detect an immediate impact of the punishment opportunity on contributions at the switch points between the punishment and the no-punishment condition. Remember that this is exactly what we observe. The introduction (elimination) of the punishment opportunity leads to an immediate rise (fall) in contributions.*"

Fehr and Gächter conduct a very similar set of experiments in [58] and the results obtained with regards to both parts of the experiment correlate with the results obtained in [57]. However, in [58], the authors specifically mention anger as the emotion that motivates punishment from cooperators. There appears to be a notion of fairness that human beings are emotionally sensitive towards since in both sets of experiments participants punished free-riders more when their contribution fell further away from the average group contribution to the public good. In accordance with this, participants reported that they felt greater anger at free-riders when the average group contribution to the public good was high and less anger when the average was lower. In addition to this, participants in both [58] and [57] expected other group members to be more angry with them if they engaged in free-riding whilst average group contribution to the public good was higher.

2.4.1.2 Positive Reciprocity

Whilst anger may indeed provoke punishment in one-shot and multiple interaction versions of the Prisoner's Dilemma game, an explanation for why positive reciprocity occurs is required. This issue is tackled by Bartlett and DeSteno in [9] who propose that the emotion of gratitude acts functionally to encourage individuals to repay favours even if such behaviour proves costly in the short-term. In [9], the authors provide evidence

from external sources such as [126] to support the hypothesis that people who experience gratitude frequently are more likely to engage in pro-social behaviours such as providing favours and volunteering time to help others. In their experiments, Bartlett and DeSteno set-up three studies: the first aims to demonstrate that gratitude has a direct effect on helping behaviour that is costly to the individual. In this study, experimental confederates provide the participant with a favour and then ask for help afterwards. The results obtained show that there is indeed a positive correlation between the intensity of gratitude felt and time spent helping others.

The second study aims to provide evidence to counter the argument that those who participated in the first study may have been more grateful and more helpful towards the experimental confederate due to an awareness of the social norms of gratitude (participants received a favour so as to elicit gratitude from them). To show that gratitude is the driver of pro-social behaviour rather than awareness of pro-social norms, participants from the first study were asked for help by both the experimental confederate (who still performs a favour for the participant) and a complete stranger. The results show that all participants in this study help both the confederate and the complete stranger when asked to. As explained by Bartlett and DeSteno, the reciprocity norm cannot explain the help offered to complete strangers.

The third study aims to provide experimental evidence that the help provided to strangers in the second study was not due to a spill-over effect from the gratitude felt towards the confederate. Therefore, in this study, the experimental confederate still provides a favour to the participant but only strangers request help from the participants. In one version of this study, participants were asked a question designed to draw attention to the confederate's favour, this version is referred to as the gratitude-source version. The standard version of the study does not draw attention to the confederate's favour. The results revealed that gratitude was most intense amongst participants in the gratitude-source version of the study but participants in the standard version of the study helped the stranger significantly more. Given these results, Bartlett and DeSteno argue that the hypothesis pertaining to gratitude promoting pro-social behaviour rather than pro-social norms is fortified. It is also argued that, by drawing the attention of participants to the pro-social behaviour of the confederate, then pro-social norms such as "pay it forward" are made salient. Consequently, there should be an increase in pro-social behaviour generally, as proposed by Reno et al. in [156]. The results however show that this is not the case and participants from the gratitude-source version of the study spend significantly less time helping a stranger than those participants from the standard version of the study.

DeSteno et al. extend their argument for gratitude being the champion of pro-social behaviour in [43]. In this investigation, economic games are used instead of requests for assistance since economic games pit potential immediate gains against immediate losses whereas requests for assistance only impact upon the time of individuals; there are usually no immediate, tangible gains resulting from such behaviour. Further to

this, the authors cite Vohs et al. [208] as support for the argument that the presence and consideration of money in such games decreases feelings of interdependence and pro-social helping. Ergo, testing for pro-social behaviour as a result of gratitude in such games can potentially provide strong evidence that this emotion is responsible for cooperation between individuals in human society. DeSteno et al. propose that gratitude increases the likelihood of cooperation even when real money is at stake. The experimental set-up for this paper is practically identical as that previously described in Bartlett and DeSteno's previous work [9]. Instead of an experiment confederate asking for help however, half the participants play a public goods game against the confederate and half against a stranger. The decisions and actions of each player were made in private and the games were of a one-shot nature. Therefore, the impact of strategic considerations upon potential long-term gains/losses were non-existent in the games i.e. any gratitude displayed was not motivated by impure altruism (helping in anticipation of some long-term pay-off). The results from [43] show that participants who reported feeling gratitude towards the confederate also contributed more on average to both the stranger and the confederate. What is especially important about these results is that they show that acts of cooperation are not brought about by consideration/awareness of pro-social norms since participants cooperated with strangers as much as they did with the known confederate.

2.4.1.3 Upstream Reciprocity

Nowak and Roch in [136] outline how positive upstream reciprocity is motivated by gratitude. Upstream reciprocity is defined by the authors as behaviour where an individual, x, helps another individual, y, because x received help from another individual, z. In this work Nowak and Roch are essentially bolstering the findings made in [43] (or laying down their foundations since, chronologically, Nowak and Roch's paper was written before DeSteno et al.'s) by using random-walks to calculate whether cooperation will survive in a population. In these games, a simulated player has two parameters: p and q; p is the probability that a player will pass on cooperation when it is received from another and q is the probability that a player will spontaneously cooperate. Therefore, each player, S, can be defined as $S(p, q)$ and four strategies/player-types are defined, these are:

- $S(0, 0)$ - Defectors. Players of this type never spontaneously cooperate, neither do they pass on cooperation received from others.
- $S(0, 1)$ - Classical cooperators. Players of this type spontaneously cooperate but do not pass on cooperation from others.
- $S(1, 0)$ - Players of this type never spontaneously cooperate but always pass on cooperation from others.

- $S(1,1)$ - Players of this type spontaneously cooperate and always pass on cooperation from others.

In order to promote cooperation, players incur a cost, c, to themselves which is less than the benefit, b provided to the recipient of the cooperation therefore $b > c$. For games that only allow upstream reciprocity Nowak and Roch demonstrate that natural selection always reduces cooperation rates. This is due to the fact that strategies with less cooperation always out-compete highly cooperative strategies and pure defectors always dominate. They therefore conclude that upstream reciprocity alone does not enable the evolution of cooperation within a population.

However, Nowak and Roch note that the outcome changes dramatically if players are allowed to reciprocate directly i.e. if they are allowed to cooperate with a player that has just cooperated with them or if players are allowed to randomly compare their current pay-off to the pay-off of a neighbour and change their strategy if it is found that this neighbour has a higher pay-off values (spatial reciprocity). If direct reciprocity allows cooperation then upstream reciprocity evolves too and, in models that include spatial reciprocity, cooperation evolves much more easily with upstream reciprocity than without. In both types of model, Nowak and Roth report that classical cooperators are always out-competed by strategies whose $p > 0$ and upstream reciprocity greatly enhances the level of altruism in a population.

Overall, their study proposes that gratitude and other positive emotions may be the key to understanding phenomena such as upstream reciprocity and such emotions may evolve through the process of natural selection.

2.4.2 Reciprocal Altruism

In [199], Trivers outlines and discusses the concept of reciprocal altruism which he argues has emerged from a process of evolution. Reciprocal altruism is behaviour defined as the following: suppose there is a man drowning in a body of water. If I am passing by I could either choose to help the man or to ignore him. Trivers assumes that the cost to me, with respect to utility, of helping the man is marginal compared to the utility that would be bestowed upon the man if I helped. If my behaviour is truly reciprocally altruistic I will help the man based upon the expectation that I will, at some point, be compensated for my action at a later date. The most interesting and relevant part of this work are Triver's predictions and discussions of the psychological system that may underlie and motivate reciprocal altruism in human beings. He posits that social constructs such as friendship are based upon positive emotions such as liking and disliking and these emotions will have evolved in order to mediate altruism in human beings. He notes that research by Sawyer [165] and Friedrichs [66] also argues that liking and friendship are drivers of altruistic behaviour. In [165], Sawyer presents results that show how altruistic behaviour is displayed more towards friends than neutral individuals whilst in [66], Friedrichs shows that more altruistic behaviour is displayed towards friends who are more attractive. It is also noted that the relationship between altruism and liking is

reciprocal as those who are altruistic display such behaviour significantly more to those one likes and altruistic people like those who are also altruistic [19], [111]. However, if such a system exists then there will be a selection pressure exerted against it whereby cheaters (defectors) may begin to take advantage of such positive emotions. In such a situation, Trivers argues that moralistic aggression will be selected for in order to ensure:

- That an altruist does not continue to be altruistic towards someone who does not reciprocate.
- Reciprocation of altruism from a non-reciprocator by threatening physical punishment or the suspension of altruism towards them.
- That in extreme cases of non-reciprocation the non-reciprocator will be removed either by injury, death or exile.

Moralistic undertones to such acts of aggression are commonplace and again Trivers points to a number of pieces of research that validate this idea. For example: Thomas [198] and Marshall [122] demonstrate that, in hunter-gatherer societies, real or even imagined injustices such as unequal food-sharing are the source of much aggression. Furthermore, it is hypothesised that gratitude, sympathy and shame have also evolved to regulate and respond to the "cost to self/benefit to other" trade-off. Trivers suggests that gratitude has been selected for in order to regulate the amount of reciprocation employed by the receiver of an altruistic act and sympathy has been selected for in order to motivate altruistic behaviour and determine how altruistic a person should be. Simply put: if an individual, x, feels more intense sympathy for another, y, then x will display increased levels of altruism towards y. Similarly if y feels more intense gratitude in response to x's altruistic action then y will reciprocate with a more altruistic act.

Again, numerous pieces of research support such psychological considerations of cost/benefit ratios. Gouldner proposes in [77] that the greater the need for an altruistic act, the greater the tendency for the recipient to reciprocate if altruism is displayed toward him/her. Gouldner also suggests that the less resources the donor of the altruistic act has, the greater the reciprocation from the receiver. In [84], Heider postulates that the intensity of gratitude experienced is greater when the altruistic act benefits the recipient. Tesser et al. [196] present evidence that American undergraduates think they will feel more intense gratitude if they are the recipient of an altruistic act that is valuable to them and costs the donor a great amount (Pruitt also offers research that augments this idea, see [153]). Additionally, a considerable literature review by Aronfreed [6] substantiates that sympathy, along with gratitude, is also a driver of altruistic behaviour in the way proposed by Trivers in [199].

In response to moralistic aggression, those who feel guilty in some way would also be selected for in response to their defection being brought to light. Trivers proposes that guilt has been selected for in humans since this emotion motivates behaviour which ensures that both the cheated and the cheater do not suffer. The reasoning for this

is such: the guilty cheater can take steps to rectify his wrong-doings and this may discourage the cheated individual from cutting off all support to the cheater. The suspension of retaliation may also benefit the cheated as the cheater may then reciprocate further as the cheated exhibits altruism by not retaliating (and not recovering their losses plus extra). However, Trivers points out that agreement on guilt being the motivating psychological factor behind reciprocal altruism is sparse. Further psychological systems that may have evolved in order to enable reciprocal altruism in human beings are also noted in this work but they are not of interest to the context of this thesis and are therefore not discussed further.

2.4.3 Other Notable Strategies

Reciprocity and reciprocal altruism can both be seen as examples of "tit-for-tat" or "TFT" behaviour, a strategy that was noted as being particularly effective in Axelrod's *Evolution of Cooperation* [7]; perhaps one of the most famous and extensive pieces of research regarding how cooperation may evolve in the iterated version of the Prisoner's Dilemma game. Axelrod's measure of success in these experiments was the total number of points earned by each strategy after the conclusion of a game (based upon the classic points system for the game outlined in section 2.1.2). In the first tournament held by Axelrod the TFT strategy emerged as superior. The strategy cooperates on the initial round and echoes the opponent's strategy in the previous round for subsequent rounds. For example: in round 2 the strategy plays what its opponent played in round 1, for round 3 the strategy plays what its opponent played in round 2 etc. A second tournament was then run again by Axelrod that solicited strategies from a wider range of sources; TFT triumphed again. Axelrod's subsequent dissection of the strategy postulated four rules that were crucial determinants of TFT's success:

1. Do not defect initially. Whilst this may bestow the highest utility upon the defector if the opponent cooperates, this small increase in utility in the short-term is offset by the danger of DD being played for the remainder of the game.

2. Do not try to "beat" your opponent, instead try to elicit cooperation.

3. Reciprocate defection/cooperation immediately with punishment/reward. TFT did not suffer defection lightly, neither did it overreact to defection. TFT was forgiving i.e. even though an opponent may have defected previously for several rounds, if cooperation was exhibited, the strategy reciprocated thus re-establishing CC. TFT did not also try and take advantage of others by periodically defecting after establishing CC with an opponent.

4. Don't be too clever. Sophisticated strategies did not take into account that, whilst it may be watching the opponent and learning, the opponent may be doing the same and no meta-learning of this reciprocal learning was exhibited. Complex strategies that were submitted were also unforgiving, resulting in cooperation not

being able to be re-established if an opponent defected in any round. Complex entries also exhibited behaviour that could have been considered random by other strategies so such behaviour did not confer any notable advantages.

Other strategies have been proposed that may outperform TFT in populations of individuals playing the iterated Prisoner's Dilemma. The first of these deserving mention is the *Generous Tit-for-Tat* strategy or "GTFT" investigated by Nowak and Sigmund in [137]. GTFT essentially delays defection so, if an opponent defects in context of the Prisoner's Dilemma game, a player who employs GTFT will not immediately defect like TFT. Instead GTFT requires *two* defections before responding with defection.

In a population of heterogeneous strategies that included TFT, where the success of a strategy is rewarded by more offspring, GTFT always emerges as the most successful. Nowak and Sigmund note however, that without the presence of TFT in a population, hardcore defectors out-compete GTFT since GTFT feeds upon TFT players. GTFT is able to do this since it has the ability to correct mistakes when playing against TFT players. For example: if a TFT based strategy defects without provocation in round n, GTFT will re-establish CC in round $n+1$. Crucially, the forgiveness displayed by GTFT imparts benefits when strategies are uncertain.

The second notable strategy that outperforms TFT with respect to cooperation rates is the *Pavlov* strategy, again notably investigated by Nowak and Sigmund in [134]. Pavlov is capable of outperforming TFT in populations comprised of heterogeneous strategies and where mutation and selection are possible. Pavlov is based around two simple principles:

1. The "win-stay" principle: if a Pavlov player, i, and its opponent, j, both cooperate in round n then i should cooperate again in round $n+1$. If, however, i defects and j cooperates in round n then i should defect again in round $n+1$.

2. The "lose-shift" principle: if a Pavlov player, i and its opponent, j, both defect in round n then both should cooperate in round $n+1$. If, however, i cooperates and j defects in round n then i should defect in round $n+1$.

As noted by Nowak and Sigmund, these principles allow Pavlov players to correct occasional mistakes such as when an unprovoked defection between two previously cooperative TFT players causes subsequent rounds to be composed of unending, asynchronous defections. In this case, there is only one round of DD for a Pavlov player before CC is restored (if both players are using the Pavlov strategy). Furthermore, Pavlov players are capable of exploiting unconditional cooperators who may undermine populations by allowing unconditional defectors to thrive.

To the best of my knowledge, no existing literature investigates the motivations behind human beings employing the strategies mentioned above therefore, there is considerable scope for analysis regarding why human beings may adopt the GTFT or Pavlov strategies. It could be that some human beings are more *tolerant* (willing to cooperate

in the face of defection) and *responsive* (ready to cooperate with an opponent after a period of defections from them) which would explain GTFT behaviour in human societies. With respect to those who may utilise the Pavlov strategy, it may be that these individuals are driven by positive emotional states such as satisfaction when winning and negative emotional states such as self-reproach when they lose.

2.5 Emotion and Cooperative Behaviour

From the discussions in section 2.4 it would be reasonable to propose that emotion is one of the key determinants of intentional behaviour in human beings especially in the context of public goods games. Many books have been written to address the relationship between cooperative behaviour and human emotion and I would direct those interested towards the most notable of these, namely *Genetic and Cultural Evolution of Cooperation* by Fessler and Haley [60]. This section now aims to provide a more general review of literature concerned with the role of emotion in causing intentional behaviour that both establishes and maintains cooperation. I conclude with a review of literature that makes the case that emotions should be modelled computationally and their effects upon decision-making studied in the context of public goods games studied since the self-interested, rational model of agency clearly falls short in explaining the behaviour of human beings in an economic context.

From the discussions presented in sections 2.2, 2.3, and 2.4, it appears that human beings do not only take into account rational self-interest when sharing resources. Considerations of fairness are included in decisions regarding others and unfairness often results in punishment, even when some cost is incurred by the punisher. Whilst not explicitly mentioned in section 2.2.1, I would posit that Ache tribe members take into account the amount of time spent helping, effort expended, risk of injury etc. in some calculation of fairness in order to determine who food should be shared with and how much should be offered. Furthermore, I would propose that the motivation to follow through on such considerations is emotional in nature i.e. gratitude may be a motivating factor as could the fear of anger being directed towards the individual in question. Literature regarding considerations of values other than selfish utility-maximisation in deciding what to do are abound, especially within the realms of argumentation where Bench-Capon proposes value-based argumentation frameworks (VAFs) [17]. VAFs are a mechanism for argument evaluation that take into account premises such as social values that are applicable to the audience that an argument is directed towards. Therefore, a different audience may come to a different conclusion regarding the acceptability of an argument if they order some particular values differently. Furthermore, Bench-Capon et al. also use VAFs in [16] to evaluate arguments that serve to explain the results obtained in Dictator and Ultimatum games such as those discussed in section 2.2.2. By modelling these games in such a way, the social values of players are accounted for and results obtained in practice are reflected. If Dictator and Ultimatum games are modelled

using the assumptions of rational game-theoretic principles however, the results do not correlate with reality. By extending these VAFs further, it is possible to include explicit emotional premises as proposed by Nawwab et al. in [131].

The role of emotions in guaranteeing threats and promises in social interactions is also discussed in great detail by Hirshleifer in [90]. In this work, Hirshleifer considers a number of one-shot games and proposes two different scaled classes of emotion that serve as threats/promises which elicit cooperation from self-interested individuals: the benevolence/malevolence class, which is action-independent, and the anger/gratitude class, which is action-dependent. These classes are deserving of further explanation; the action-dependent class takes into account emotions that enable a person to act passionately or as Hirshleifer defines the concept:

> "what he [a person] wants to do need not depend only upon the final outcome in the utilitarian sense - i.e. strictly upon the ultimate distribution of incomes between the two parties - but rather may be **action dependent**."

Essentially Hirshleifer proposes that in some situations, expected utility is no longer considered in determination of intentional behaviour and instead, the perceived intentional behaviour of an agent, x, may cause the elicitation of anger/gratitude in another agent y, which motivates y to perform some intentional behaviour (cooperation or defection). Hirshleifer also argues that, in certain circumstances, such non-utilitarian behaviour makes complete utilitarian sense. It is proposed that preference orderings on incomes are not appropriate in such circumstances as it is the actions of the so-called "first-mover" (agent x) that are of interest to the "second-mover" (player y) rather than the utility entailed by those actions. The intentional behaviour demonstrated by y is therefore dependent upon the emotion elicited (and the intentional behaviour exhibited by x). For example: the more grateful y is, the more likely it is that y will confer benefit upon x. Similarly, the more angry y is, the more likely it is that y will confer injury upon x. This can produce the intentional behaviour detailed in sections 2.2, 2.3 and 2.4 since, whilst x may increase its wealth at the expense of y, y can then punish in return (even when there is a cost to y) due to the emotion elicited overriding utilitarian consideration. This moves both players towards a mutually beneficial outcome as x will attempt to reduce y's anger by conferring a benefit upon it provoking cooperation from y when gratitude is elicited (enlarging x's wealth once again). Thus, the role of anger and gratitude can be thought of as being guarantors of threats and promises or cooperation and defection. This is a function of emotion which I will now turn attention towards.

The use of emotion or conscience as a commitment device to ensure cooperation has also been noted by Frank in [63]. The ideas presented in this work build upon those presented by Trivers in [199] as it is proposed that people who commit to certain actions but then do not follow through with them may experience the emotion of guilt. If this feeling of guilt is sufficiently strong it may then prescribe actions from the guilty party that enable cooperation even when such actions do not promote the self-interest

of the individual in question. In related work, [62], Frank also demonstrates how guilt is capable of sustaining cooperation in the Prisoner's Dilemma game. Since guilt entails a generally negative and unpleasant feeling, individuals are motivated to avoid eliciting it and will therefore avoid actions that trigger such an emotion. Instead, guilty individuals will engage in cooperative actions that are likely to increase the positivity of their mood (see [33] and [30]), even for some time after the guilt-eliciting situation has passed [104].

Further to the discussions provided in sections 2.4.1 and 2.4.2, the link between emotion and cooperative behaviour has been studied quite extensively. For example: Aderman [3] provides experimental evidence supporting the notion that individuals experiencing positive emotional states are more likely to volunteer for unpleasant future scenarios than participants who are currently experiencing negative emotions. Isen provides further experimental evidence which substantiates the proposition that people who succeed at a particular task (therefore inducing a positive mood) are more likely to behave generously and cooperatively towards a complete stranger than those who did not succeed in the same task [92]. The amount of literature regarding the issue of positive emotional states increasing cooperation rates is quite extensive, with works such as [61], [93] and [129] all contributing experimental evidence. Work by Pilluta and Murnighan [148] pits two models of explanation for the rejection of Ultimatum game offers against each other: an "unfairness" model and the "wounded pride/spite" model developed by Straub and Murnighan [191]. The "wounded pride/spite" model asserts that respondents who are fully informed with respect to the total amount of money to be divided and the actual division proposed in an Ultimatum game may feel emotions i.e. anger, at unfair offers resulting in the rejection of offers that are economically valuable. The results obtained demonstrate that the emotional model proposed is capable of explaining the behaviour observed in their investigation and that anger is indeed an important determinant of people's decision-making in an Ultimatum game context. Previously mentioned research by Straub and Murnighan [191] also shows that responders who are made small but positive offers under incomplete information circumstances reject these offers. This is because responders feel that small amounts of money are not worth considering, a result that was also verified in separate research by Murnighan and Saxon [130].

The research discussed above again provides evidence that the *homo economicus* model of agency does not hold and that people take more than just utility maximisation into account when deciding what to do next. With respect to these other considerations, Bosman and Winden [21] outline the concept of "emotional hazard" and assert that human beings consider this when deciding upon whether to foster cooperation or destroy it. "Emotional hazard" is defined as how an offer of utility may affect a particular emotion and in turn may cause a destruction of the utility offered. Therefore, emotional hazard should be taken into consideration when making a decision about how much to offer another, especially if they are capable of destroying your own potential utility (as in the Ultimatum game). Bosman and Winden use simple "power-to-take" games in [21]

to investigate how emotions generate behaviour. Participants first have to earn some income individually, and are then randomly paired with one participant becoming the taker and the other becoming the responder (as in the Ultimatum game); the game then proceeds in two stages. First, the taker proposes an amount that will be taken from the responder; second, the responder may decide how much of its income to destroy. The following results were observed:

- The behaviour of responders is discontinuous, they either completely destroy their income or leave it untouched.
- The intensity of negative emotions felt by the responder increases as value of proposals increase. Conversely, the intensity of positive emotions felt by the responder increases as the value of proposals decrease.
- The probability of the responder destroying income increases as the intensity of negative emotions increases.
- The probability of the responder destroying income increases as the value of proposals increase.
- The responder's expected value of the proposal has a significant effect on destroying income but not on the intensity of emotions experienced.

Bosman and Winden also use a modified version of the game detailed above to investigate other questions. The no-effort experiment simply endows participants with income rather than them having to earn it. There are some notable differences in the results observed between the effort-required and no-effort experiments. The only result I will discuss here (since it is the only one relevant to this thesis) is that in the no-effort experiments, responders destroy an intermediate amount of income more frequently as well as destroying a greater aggregate amount of income. A further two results are related to this:

- Responders who are observed to have destroyed their entire income are significantly more irritated than those who destroyed lesser amounts.
- The same emotional intensity is experienced by responders who destroy all their income; effort has no effect.

The results are interesting as they lend support to the idea proposed by Hirshleifer in [90] i.e. when an emotion is sufficiently intense (the intensity has reached some threshold) the punishment of an unreasonable action takes precedence over utilitarian considerations. At high intensities, Bosman and Winden note that all income is destroyed as punishment even though this adversely affects the punisher. In response to this result, the authors propose that:

> "...at higher emotional intensity individuals do not make a compromise between their conflicting intrapersonal urges (here, going for the money versus punishment)."

The notion of *emotional thresholds* are alluded to in the experiments conducted by Bosman and Winden in [21] since participants tend to destroy either all of their income or nothing (at least in the effort-required experiments). Therefore, there would appear to be some sort of "tipping point" for emotion where, if the emotion is not intense enough, no intentional behaviour will be exhibited as a consequence of that emotion being elicited. Furthermore, the existence of an emotional state in an individual does have the potential to shift value ordering and preferences since behaviour that promotes utilitarian values is demoted in favour of values that promote emotional considerations (the destruction of all income rather than none or some, for example). In addition, the results from both Bosman and Winden's effort-required and no-effort experiments illustrates the direct relationship between emotional intensity and non-utilitarian behaviour. Finally, results from the aforementioned experiments also highlight the relationship between the intentional behaviour of others and our own emotions. This is demonstrated most clearly when participants report higher intensities of positive/negative emotion as the proposed take rate becomes increasingly fair/unfair [21].

2.6 Chapter Summary

In this chapter I attempted to consider a broad survey of literature regarding social interactions between human beings in the context of public goods games. By doing so, my intention is to elucidate the differences between *homo economicus* and *homo sapiens* with respect to observed behaviour in experimental settings and to explain why these differences exist. To achieve these goals, I first considered the self-defeating behaviour predicted by the *homo economicus* agency model in context of well known public goods games such as the Ultimatum and Prisoner's Dilemma games in section 2.1. This prediction of behaviour was then contrasted in section 2.2 with behaviour observed in a broad survey of anthropological and economic literature concerned with the investigation of human behaviour in public goods game environments. Section 2.3 further honed my considerations by reviewing literature that refers to the behaviour of human players in Prisoner's Dilemma games. The research taken into account by both sections 2.2 and 2.3 illustrates how the bleak predictions made by the *homo economicus* model are not generally adhered to by human society.

After highlighting this difference I shift focus onto providing an explanation for why this discord between the *homo economicus* model of agency and the behaviour of human players in public goods games such as the Prisoner's Dilemma exists. In sections 2.4 and 2.5, I posit that *emotion* is the integral reason for this divide following a consideration of notable strategies that promote cooperation. By affording emotion a role in decision-making in public goods game contexts, the behaviour exhibited by human players can be

better explained than by using the narrow *homo economicus* model of agency proposed by game theory.

The key points to take from this chapter are listed below. In raising these issues I aim to clarify a research agenda that may be embarked upon in the remainder of the thesis.

- The decision-making and cooperation rates exhibited by human participants both in reality and in games such as the Ultimatum and Dictator games, is markedly different from the predictions made by the *homo economicus* model of agency. This dissonance is most noticeable if the behaviour of human participants whilst playing Prisoner's Dilemma games is exclusively considered.

- The utility maximised by some players is not just income. "Utility" could also refer to a person's emotional state, their perception of fairness, their social standing etc. It is proposed that emotion causes people to alter the values they consider when it comes to making a decision about whether or not to cooperate or defect with others. Whilst some game theorists argue that "utility" can be taken to mean any number of values besides wealth, such reasoning becomes implicit due to the heavy focus given to monetary gain as demonstrated by the *homo economicus* model of agency.

- Emotions facilitate decision-making when one must decide how to act in context of unpredictable environments where complete information is not able to be acquired.

- Emotional states such as anger and gratitude appear to be taken into consideration by human players in public goods games when deciding upon what actions to take. Clear correlations between behaviour exhibited and emotions reported (whether arising from pre-game conversations or during the game) have been identified by many researchers (see [21, 56, 58, 62]).

The idea of incorporating emotion into decision-making has been argued for by economists such as Lowenstein [120] and Elster [50]. With particular reference to [120], Lowenstein outlines his point of view that visceral factors should be included in economic decision models since they serve essential functions. Visceral factors are characterised by the immediacy of their effects i.e. if one experiences happiness in response to helping another then one would wish to "re-do" that action in order to feel happy again. The effects of visceral factors are important as they are not as transient as typically thought by economists and neither are they too complicated/unpredictable to be formally modelled. Lowenstein also states that economic models that ignore visceral factors only really begin to approach predictive accuracy when the behaviour described can be characterized as being relatively immune to the existence of visceral factors.

The question that stems from this chapter is quite philosophical in nature: if the effects of emotions upon intentional behaviour is to be modelled computationally for use in agent systems is such intentional behaviour *part* of an emotional experience?

The question of whether emotions should even be considered for use in computational systems should also be asked since it has long been thought that the presence of emotion in decision-making reduces the perceived "rationality" of the decision made. These questions are the focus of chapter 3.

Chapter 3

Emotion, Behaviour and Rationality

This chapter considers philosophical arguments related to the concept of "emotion" pertinent to this thesis. I will, in particular, consider the relations between emotions and intentional behaviour. For example: it may be questioned as to whether an emotion such as fear can be said to have been elicited in the presence of a behaviour such as "running away" but in the absence of a behaviour such as "sweating profusely". The reason why this question is so important is outlined in section 3.1 (specifically section 3.1.1) along with discussions of philosophical accounts of emotion that are intended to provide an answer to the question posed. Furthermore, section 3.1.2 is also intended to clarify my own views regarding when emotion occurs in respect to the emergence of an *intentional* behaviour. This question has been the subject of much philosophical and psychological debate for centuries and so it is important to state my position with respect to this issue as it is one of the central foundations this thesis is built upon. In these discussions the concepts of *action potentials* and *emotion thresholds* will be mentioned and are important to note since they will play a key role in the emotion modelling framework that I will advance in chapter 5.

The chapter's focus then shifts to contemplate the supposed dichotomy that exists between the concepts of "emotion" and "rationality". This issue is discussed in section 3.2 and stems from the belief and assertion by others such as Slovic et al. [177] that emotions and decision-making do not need to exist in tandem for rational decision-making to occur. In other words, emotions have no intrinsic role in the process of decision-making; they are purely adjunct. Some academics such as Martino et al. [123] extend this argument to posit that emotions should be defined as "irrational" in the sense that emotions produce irrational behaviour when employed in decision-making. Again, answering this philosophical question is an integral concern with respect to the thesis since, if emotions truly are irrational and impact upon rational decision-making negatively, then why should effort be invested in developing computational models of emotion? Computer systems are intended to make rational, logically consistent decisions

and including emotion in the decision-making of such a system could be seen as being detrimental rather than beneficial. In this chapter and specifically in section 3.2, I hope to illustrate why this is not the case.

3.1 Emotion and Behaviour

In chapter 2, it was proposed that human behaviour in public goods game contexts appears to be driven not only by self-interested rationality but also by emotion. To construct a functional, computational correlate of emotion that enables a software agent to decide upon what behaviour to employ (so that the effects of various emotions upon social interactions in public goods games can be analysed), I will need to first ascertain whether or not an emotion can be said to exist in the absence of unintentional behaviour (see chapter 2's introduction for a definition of unintentional behaviour in the context of this thesis). With respect to this issue, it is important to note that a computer is incapable of producing unintentional behaviour for two reasons:

1. A computer does not possess the necessary physiological mechanisms by which to produce such behaviour.

2. Since a computer operates by virtue of calculations and programs every action performed by one is intentional (whether the intention originates from the programmer or the machine: as Freud states: "*nothing is accidental*"). Consequently, genuine unintentional behaviour cannot be produced in such a machine.

If it is the case that unintentional behaviour is a part of emotion and cannot be separated from it then I assert that a software agent cannot experience a proper emotional state and if this is so, I can only *simulate* rather than *create* emotions. Being clear on this point will ensure that I do not address incomplete emotions as complete and vice-versa. To this end, I discuss the matter in section 3.1.1 and propose an answer to the question of whether or not there is a direct relationship between intentional behaviour and emotion and can it be asserted that an agent experiences an emotion in the absence of unintentional behaviour?

In addition to answering this question I also consider whether or not emotion can truly motivate intentional behaviour or whether an emotion occurs after intentional behaviour has been expressed. Whilst the former statement may seem commonsensical, this issue has been (and is still) debated fiercely by psychologists and philosophers alike. Some experts even go as far to propose that emotions only have a direct impact upon cognition and have no immediate influence upon intentional behaviour. The second half of this section i.e. section 3.1.2, seeks to clarify this issue by considering such arguments to answer the question of how intentional behaviour is related to emotion.

3.1.1 Emotion and Intentionality

Being able to assert that unintentional behaviour is not necessarily part of an emotional experience and that an emotion can exist in the absence of such behaviour but in the presence of intentional behaviour, is very important if emotions are to be ascribed to computers, in any sense. Unfortunately, philosophies and models of emotion are fraught with opposing views with respect to whether unintentional behaviour is a part of emotion. Consequently, the following section considers a number of philosophical theories of emotion so that I may develop a valid, well-supported and accurate mode of thinking with respect to this issue.

3.1.1.1 Spinoza's Philosophy of Emotion

Within Spinoza's "*The Ethics*" [185], two chapters are concerned with emotion; the first regards the origin and nature of emotions whilst the second is concerned with how emotions "enslave" the human mind. Spinoza conjectures that the mind and body are both part of nature and are therefore both subject to its laws. Consequently, this implies that Spinoza considers all behaviour resulting from emotion to be unintentional as the mind will never be able to act autonomously to produce intentional behaviour. Spinoza believes that perceptions and appraisals of external events give rise to emotions and that unintentional behaviour is a part of emotion rather than a consequence of it. Since gratitude has already been discussed in sections 2.4.1 and 2.4.2, I will consider Spinoza's treatments of this emotion here to clarify his philosophy.

According to "The Ethics", gratitude is defined as:

> "the desire or zeal springing from love, whereby we endeavour to benefit him, who with similar feelings of love has conferred a benefit on us".

Thus, according to this definition, if "we" is denoted as x and "him" as y then gratitude has two premises. The first of these premises is that y must perform an action that is beneficial towards x. The second premise is that love must exist between both x and y since y's motivation to confer a benefit on x and x's motivation to reciprocate such a benefit are both due to the existence of love between these agents. Thus, for Spinoza, gratitude only arises in the context of mutual love, rather than giving rise to love, as is often thought. The definition prescribes that the elicitation of gratitude in x produces an unintentional behaviour that confers a benefit upon y due to some beneficial behaviour being received. Therefore, the unintentional behaviour is a part of the emotion rather than a consequence of it. It should also be noted that Spinoza explicitly states that gratitude is direct in its effect i.e. x endeavours to benefit y (an example of positive reciprocity, see chapter 2, section 2.4.1.2) rather than some other agent (an example of upstream reciprocity, see chapter 2, section 2.4.1.3).

For love to be elicited, a further two premises must also hold according to Spinoza's definition of the emotion: an agent must first experience pleasure and second, this

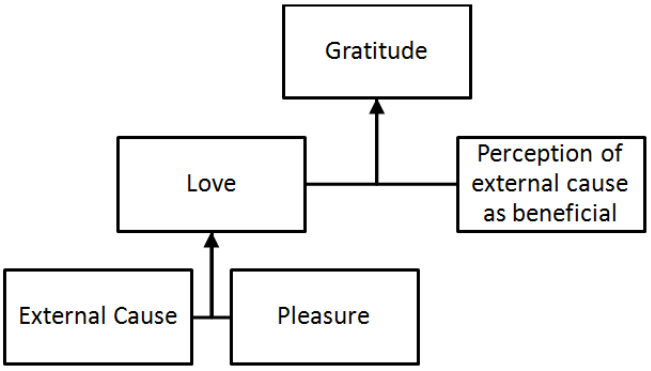

FIGURE 3.1: The premises of gratitude according to Spinoza.

pleasure must have been produced by some external cause. Therefore, according to Spinoza, the root of gratitude is pleasure, defined as: "*the transition of a man from a less to a greater perfection*". This implies that some appraisal needs to be undertaken as the transition from a less to a greater perfection needs to be recognised. Based upon these definitions a tree of premises and consequences may be constructed that creates a fairly tractable model of emotion (see figure 3.1).

Considering the discussion thus far, it would appear that Spinoza proposes emotions to be composed of two factors: the perception/appraisal of an external event and an unintentional behaviour (this is similar to Descartes whose philosophy of emotion will be discussed in section 3.1.1.2). Whilst both philosophers agree that unintentional behaviour is a part of emotion, Spinoza does not recognise the concept of intentional action since he does not agree with the notion that the human mind is autonomous.

The tractability of Spinoza's emotional model is beneficial from a computer science standpoint since the tree of premises shown in figure 3.1 can be easily translated into program instructions. Modelling emotions in this fashion would allow any given emotion to be modelled in much the same way and is used notably utilised by Ortony et al. in their appraisal model of emotion [142] (discussed later in chapter 4) and is an important idea with respect to this thesis.

3.1.1.2 Descartes' Philosophy of Emotion

Descartes' "*Passions of the Soul*" [42] is an attempt by the philosopher to outline his philosophy of the passions or emotions. In this work, a "passion" is defined as:

> "those perceptions, sensations or emotions of the soul which we refer particularly to it, and which are caused, maintained and strengthened by some movement of the spirits"

To put this in more specific terms, a passion is a mental state that arises from brain, nerve and other physiological activity. This proposition means that Descartes work is one of the earliest investigations which connects knowledge of human physiology to emotion, allowing his arguments to be grounded in physical evidence rather than abstract thought

alone. The "spirits" that are referred to in this definition warrant some explanation as such a term may nowadays imply that Descartes is invoking the paranormal to explain how passions/emotions are produced. "Spirits" are defined in the text as "*lively parts of blood that are distilled by the brain and heart*", in modern physiological terminology Descartes is referring to what is now understood to be nerve impulses, connections between neurones in the brain, transmissions across synapses etc. It is proposed by Descartes that these impulses create unintentional physiological behaviours in the body and prepare the mind to potentially undertake some intentional behaviour as a result of the emotion elicited.

The major difference between Descartes' and Spinoza's philosophies (see section 3.1.1.1 for a discussion of Spinoza's philosophy) is rooted in this recognition of a relationship between intentional behaviour and emotion. Unlike Spinoza, Descartes asserts that the mind and body are separate, distinct entities and whereas Spinoza believes that all behaviour is unintentional, Descartes rejects this idea. Instead, Descartes asserts that human beings may gain control over their emotions, meaning that some kind of intentional behaviour may be produced by the experience of an emotional state.

The implied division between emotion and intentional behaviour is important as Descartes believes that intentional behaviour is not part of an emotion; it is a consequence of one. Essentially, Descartes recognises a functional link between emotion and intentional/unintentional behaviour and identifies a number of passions/emotions and dissects each in some detail. I will consider Descartes' analysis of gratitude for the same reason outlined in the discussion of Spinoza's philosophy of emotion in section 3.1.1.1. By doing this I also hope to make the distinctions between Descartes' and Spinoza's philosophies of emotion clear.

Descartes asserts that gratitude is elicited in a person, x when the action of another, y, has attempted to do some good for x. It is posited that the experience of gratitude by x causes it to *want* to reciprocate the action[1] which stands in contrast to Spinoza who posits that gratitude *always* causes reciprocation.

What Descartes appears to be outlining is the concept of *action tendencies* which will feature prominently in chapter 4, section 4.1.1.2. In very abstract terms, Descartes proposes that the experience of an emotion causes an unintentional behaviour to be exhibited in the experiencing agent but only the potential for an intentional action is created. To clarify this, Descartes states that with respect to fear, the "spirits" always have an effect on the nerves in the legs, the expansion and contraction of openings to the heart, and the nerves that agitate other parts of the body from which blood is sent to the heart. Intentional behaviour on the other hand is indirectly influenced by the "spirits" since the necessary resources to perform an intentional behaviour are made available to the body.

[1]It is not specified as to whether Descartes believes that gratitude causes x to want to reciprocate the action to y (positive reciprocity, see chapter 2, section 2.4.1.2) or to some other agent (upstream reciprocity, see chapter 2, section 2.4.1.3).

It would appear that Descartes believes that an emotion is constituted of two parts and one consequence: the perception/appraisal of an external event and an unintentional behaviour are the two parts whilst intentional behaviour is the one consequence. This advances Spinoza's philosophy where only two parts are considered (perception/appraisal of an external event and unintentional behaviour). So, since the perception/appraisal of an external event and the unintentional behaviour in response to this is part of an emotion it can be asserted that, according to Descartes, if a perception exists without an accompanying unintentional behaviour then an emotion can not be said to exist. Such an assertion would argue that, from Descartes' point of view, emotion may never be completely represented in a computational system since unintentional behaviour is a part of an emotion and unintentional behaviour cannot be manifested in computer systems. What is notable however, is that Descartes recognises a functional link between emotion and intentional behaviour.

3.1.1.3 Hume's Philosophy of Emotion

Hume's theory of emotion is outlined in "*The Treatise of Human Nature*" [91] and follows in the appraisal-focused philosophies of emotion advocated by Spinoza and Descartes (see sections 3.1.1.1 and 3.1.1.2 respectively). Hume proposes a double relationship between "impressions" (emotions and sensations) and "ideas" (less lively copies of impressions) but, most importantly, Hume claims a link between emotion and practical reasoning making the treatise an exceptionally important piece of literature in the context of this thesis. Hume's theory is the first of all the philosophies discussed so far to propose that intentional behaviour is a part of emotion and not just an adjunct consequence.

Hume's theory of emotion is centred around the notion of impressions of which there are two types: impressions of *sense* and impressions of *reflection*. Impressions of sense appear to encompass sensations that originate from a human's internal and external environment i.e. sight, sound, perceptions of pleasure and pain etc. whilst impressions of reflection encompass all passions and sentiments i.e. pride, reproach, love, hate etc. Impressions of reflection would therefore appear to be created by reflecting upon impressions of sense i.e. I feel pain (impression of sense) so that I feel angered since I am reflecting (impression of reflection) upon this pain. Impressions of reflection may then be divided further into direct and indirect passions; Hume argues that direct passions come first from an idea (perceiving a piece of art, for example) and then an impression that is associated with this idea (pleasure, for example). Indirect passions are altogether more complex and have an association between two ideas accompanied by an association between one impression and another. Therefore, I may have some initial idea with regards to a subject and this produces an impression. So, I may perceive a piece of art (an idea) that gives rise to pleasure (an impression) and if I know that I created the piece of art (an idea) then I may then associate the pleasure currently/previously experienced with pride (an impression) as they are both positive in nature.

Whilst Hume covers in great detail how an emotion is elicited, specific behaviours in relation to particular emotions are not considered however, a link between emotion and practical reasoning is posited (as mentioned previously in this section). It is claimed that, in the absence of emotion, there is no drive to act (to perform intentional behaviour). In this sense, Hume appears to agree with Descartes (see section 3.1.1.2) in that there is a functional relationship between emotion and intentional behaviour and even goes as far to argue that only passions may select ends.

3.1.1.4 Kenny's Philosophy of Emotion

In his book "*Action, Emotion and Will*" [103], Kenny proposes that an emotion is composed of three elements: a stimulus, a physiological symptom and a deliberate, intentional response. Following on from the discussions regarding Descartes and Hume and their recognition of a functional link between intentional behaviour and emotion (see sections 3.1.1.2 and 3.1.1.3 respectively), Kenny's assertion is again vital in underpinning the functional model of emotion proposed in my work. Importantly, Kenny's philosophy of emotion proposes that, if two of the elements previously listed do not exist, then an emotion cannot be said to be experienced. But if at least two of the aforementioned elements are present, this is sufficient to assert than an emotion exists in an agent; this is especially important with respect to the thesis. Crucially, Kenny's philosophy of emotion allows for emotions to be composed of at least a perception of an external stimulus and an intentional action.

To illustrate Kenny's point, consider a soldier in some environment and the three elements mentioned previously (stimulus, unintentional behaviour and intentional behaviour) along with an emotion (fear). For the purposes of this example suppose that the stimulus is that of an attack by the enemy, the unintentional behaviour is that of profuse sweating and the intentional behaviour is to run away. This allows the truth table shown in table 3.1 to be produced.

As can be seen in table 3.1, at least two of the three components (stimulus, unintentional behaviour, intentional behaviour) must exist in order for an agent to claim the experience of a particular emotion. Kenny's philosophy is crucial to this thesis since he argues that an agent can still be said to have experienced or be experiencing an emotional state even if no unintentional behaviour was or is produced. As a result, Kenny refutes the philosophies of Spinoza and Descartes (see sections 3.1.1.1 and 3.1.1.2 respectively) in that he asserts that unintentional behaviour is *not* a part of emotion. Furthermore, Kenny also recognises the functional link between emotion and intentional behaviour that is required in order for emotion to be able to be computationally modelled in a functional context. It is however proposed by Kenny that, if unintentional behaviour is removed and only a stimulus and an intentional behaviour hold, then the associated emotion may only be *simulated*. Even if this is so, for the purposes of this thesis, being able to simulate an emotion is sufficient since a functional correlate of emotion is to be developed rather than an exact replica.

TABLE 3.1: Truth table outlining Kenny's proposed relationship between stimulus, unintentional behaviour, intentional behaviour and emotion.

Stimulus	**Unintentional Behaviour**	**Intentional Behaviour**	**Emotion**	**Description**
Y	Y	Y	Y	Fearful
Y	Y	N	Y	Controlled fear
Y	N	Y	Y	Fearful
Y	N	N	N	Fearless
N	Y	Y	Y	Irrational fear
N	Y	N	N	Medical condition?
N	N	Y	N	Emergency?
N	N	N	N	Normal

Kenny's philosophy of emotion is also beneficial in that it suggests the possibility that the same object may give rise to different behaviours and different emotions. Therefore, whilst I may run away when I see a spider another person may freeze on the spot i.e. intentional behaviour is not uniform across emotions. Rather there is a group of unintentional/intentional behaviours that have a "family resemblance" as defined by Wittgenstein in [211] (such behaviours have common features but no one feature is common to all) which may be expressed.

With respect to the stimulus of an emotion; Kenny's distinction between sensation and emotion is particularly interesting as it is proposed that an emotion has an object whilst a sensation does not. For example: if a man runs past me on the street I cannot infer anything about his emotions until I ascribe some object to his behaviour. Therefore, if I know that the man is running away from a horse I could infer that he is fearful of the horse otherwise I cannot confidently infer the emotion of fear solely based upon a perception of his behaviour (he may simply be jogging). Of course, not every stimulus-symptom-response triple relates to an emotion, but every emotion requires association with all three of these elements.

3.1.2 Emotion-Intention Timing

Having provided some philosophical backing for the view that emotions can be attributed when only stimulus and intentional behaviour are present, I now turn my attention towards establishing when emotion occurs in respect to the emergence of its associated intentional behaviour. Since the emotional model I am looking to implement has emotion as the sole motivator of intentional behaviour, evidence needs to be provided to argue that an emotion does not always occur *after* the intentional behaviour has been

performed. So, without an adequate argument capable of resisting this school of thought, I have no basis upon which to espouse emotion as the primary determinant of the intentional behaviour detailed in sections 2.2, 2.3 and 2.4.

Whilst it may seem commonsensical to argue that an emotion occurs before its associated intentional behaviour is expressed, this notion has been challenged in the past and continues to be today. Therefore, it is only right that I consider alternative theories of when emotion occurs with respect to intentional behaviour. By doing this I highlight why I model emotions in the way proposed in this thesis. This is not to say that I completely disagree with the emotional theories that are not selected to be taken forward but, since I am looking to model particular emotions, some of the theories are not applicable.

To achieve this, I consider a number of prominent psychological theories of emotion that argue for different ways of placing emotion with respect to both intentional and unintentional behaviour. Therefore, this section is split into a further two sections: section 3.1.2.1 discusses theories where emotions occur after intentional behaviour is exhibited whilst section 3.1.2.2 outlines theories that posits the occurrence of emotions before intentional behaviours.

3.1.2.1 Emotion After Intentional Behaviour

Baumeister et al.'s feedback theory of emotion ([11] and [10]) seeks to both critique the theory that emotion causes or occurs before intentional behaviour and to propose an alternative to replace it. Both works cover the same ground yet [11] is more succinct in its discussion so any references are made in context of this paper rather than [10]. The concept of feedback loops presented in [11] and [10] is plausible and could well be used to augment direct causation theory however, it appears that the authors intend for feedback theory to replace direct causation theory (the theory that emotion acts directly upon intentional behaviour) completely. I begin the discussion of Baumeister et al.'s work by first defining what the feedback theory of emotion entails before moving on to discuss their critique of direct causation theory. Highlighting and responding to the arguments relevant to this thesis is important since they may be used to attack the emotion modelling framework proposed.

Feedback theory states that, when a new object or event is perceived or experienced by an agent, the agent will perform some behaviour as a result of this perception or experience. After the object or event has been affected by this behaviour, an emotion is elicited which motivates the agent to appraise the whole episode and learn from the emotional experience. If the event or object is perceived again in the future, affective states associated with that emotion are used in order to cope with the event/object rather than the use of a "full-blown" emotion[2]. Whilst I do not disagree with the concept of feedback loops, I believe that the removal of emotion as an intermediate between perception and experience of an event or object is fallacious.

[2]This concept has parallels with Damasio's "somatic marker" hypothesis [35].

The definition and distinction of the terms "emotion" and "affective states" (Baumeister et al. cite Russell's work in [160] as inspiration behind using these terms) proposed in [11] and [10] is the foundation upon which their criticism is built. Emotions are defined as a single conscious state (maybe consisting of blends of other states to give one ultimate state), are slow to occur and to dissipate, and are usually the result of cognitive evaluations. "Affective states" on the other hand are quicker to occur, are never composed of blended mental states and require less cognitive processing to be elicited than emotions.

The salient argument put forward by Baumeister et al. is that emotions act upon cognitive processes which have an input upon the regulation of decision-making following the performance of an intentional action. Consequently, Baumeister et al. propose that emotions have an indirect effect upon intentional behaviour since the time it takes for an emotion to arise is much too long for the emotion to be considered useful in decision-making in the heat of the moment. Whilst it is not denied that emotions may have a direct effect upon intentional behaviour (which would appear to contradict their earlier argument that emotion does not act directly upon intentional behaviour), Baumeister et al. argue that such effects are sporadic and counter-productive. To further support this point, Baumeister et al. refer to accounts from people who, when placed in a situation that elicits intense fear for them, remain calm and clear headed initially but then are overcome with fear afterwards. Based upon such evidence they assert that emotion *must* occur after behaviour since the emotion is only recognised after some behaviour has been noted. There is a failure here to recognise that a person may resist the behavioural effects of an emotion if there are reasons to do so i.e. an emotion creates the potential for an action to be performed but does not guarantee this behaviour being performed. Indeed, the criticism does highlight an important point that the functional correlate of emotion I am proposing needs to account for intentional behaviour not being produced even though an emotion may be active. This issue is addressed in chapters 4 and 5 and so will not be expanded upon here.

Baumeister et al. also argue against direct causation theory by disputing the notion that specific emotions cause specific behaviours i.e. fear causes a person to flee in any circumstance, satisfaction causes the person to celebrate extravagantly in any circumstance. According to the authors: "*emotions are thus not specific enough to give rise to specific behaviours, as the direct causation theory requires*". The argument's premise is based on the assumption that those who agree with the direct causation interpretation of emotion must also agree that a specific emotion *always* triggers a specific intentional behaviour, independent of context or culture. As far as the literature reads however, direct causation theory does not assert that every human emotion produces a single, specific intentional behaviour when that emotion is activated. Ekman states that there are only specific facial behaviours for the basic emotions that may be universally recognised [45] and does not mention other types of behaviour (unintentional or intentional) for other emotions. Frijda in [67] proposes that emotions create *action tendencies* i.e.

intentional behaviours that can be tended towards depending upon the current concerns of the subject. Since the context in which these concerns arise may differ then different actions may be produced even if the same emotion is elicited in the different contexts. Cannon's second argument in [28] notes experimental evidence which indicates that visceral changes associated with one particular emotion also occurs when a different emotional state is experienced. Throwing aside direct causation theory based upon such an argument seems both over-zealous and erroneous.

It is also proposed by Baumeister et al. that, whilst it may be true that emotions evolved to directly influence intentional behaviour in other organisms:

> "...this function of emotion has been rendered somewhat obsolete by the further evolution (in human beings at least) of a complex and powerful cognitive system and a sophisticated capacity for self-regulation."

Essentially, it is posited that human beings are capable of overriding the behaviour that may be produced in response to an emotion and therefore, emotion does not directly cause intentional behaviour. Personally I consider this to be a flawed argument since, although human beings are capable of overpowering the expression of an intentional behaviour caused by an emotion, the emotion still sets up the tendency for a behaviour to be expressed (Baumeister et al. even acknowledge this by stating that emotion engenders a behavioural impulse).

Another point argued in [11] is that it is usually negative emotions that are invoked to fortify the argument that emotions have a direct effect on behaviour. It may be that the examples used typically use negative emotions such as fear, but this need not be the case. As we shall see later in chapters 6 and 7, the positive emotions of gratitude and admiration also give rise to characteristic intentional behaviours.

3.1.2.2 Emotion Before Intentional Behaviour

This section will discuss the stance taken by some psychologists that emotion occurs before intentional behaviour and exerts a direct influence on such behaviour (direct causation theory). Whilst this theory may not be entirely correct for all emotions, it is at least true for some and therefore, theories of emotion that advocate this notion should be mentioned. I begin by outlining Lowenstein and Lerner's review of the role of affect in decision making [118] before moving onto Zeelenberg et al.'s concept of "*feeling-is-for-doing*" [215].

Loewenstein and Lerner In [118], two types of emotion are proposed: "expected" and "immediate" emotions. Expected emotions are emotions that are taken into consideration when deliberating about the consequences of an intentional behaviour i.e. will a negative emotion be experienced as a consequence of some particular behaviour and if so, how intense would that emotion be? In this way, emotion can be viewed as expected utility insofar as the decision regarding what intentional behaviour to perform

is based upon how one expects to feel emotionally as a consequence of performing that intentional behaviour. Immediate emotions on the other hand are experienced at the time of decision-making. With respect to this thesis, I am interested in the immediate emotions of Lowenstein and Lerner and whilst I do not think that expected emotions are unimportant, I consider them outside the scope of this work.

As argued by Lowenstein in [117], the more intense an immediate emotion is, the less it is mediated by expected emotions or cognitive processing. Again, the concept of *activation thresholds* makes an appearance since Loewenstein and Lerner give evidence to support the idea that it is only when an emotion is highly intense that it has a direct effect upon a person's intentional behaviour. The effect of immediate emotions are also exemplified by investigations conducted by Goldberg et al. [76] where anger is intentionally induced in human participants by showing a video of a man violently attacking a helpless teenager. Following this, some participants are told that the man was punished whilst others were told that, due to a legal loop-hole, no punishment was incurred. Participants were then asked to read fictional legal cases regarding different crimes and to specify what punishments they considered appropriate for the defenders. Those who had been told that the man in the video received no punishment viewed harsher punishments as appropriate in the fictional cases. Such a result would indicate that in some cases, emotion does have an immediate effect upon behaviour.

It is also proposed by Lowenstein and Lerner in [118] that some emotions are adaptive, time-tested responses to universal, re-occurring situations. Lazarus' work in [108] is cited as support for this claim since he proposes that each emotion has a core-relational theme (a convenient summary for the harm or benefit related to the person and their current environment). If a core-relational theme is matched then the corresponding emotion is elicited and an action impulse is produced. For example: if a demeaning offence against me has occurred then anger will be elicited in response and an appropriate action impulse will be selected. Note that there is no specific behaviour attributed to the emotion elicited, which stands in opposition to Baumeister et al. (see section 3.1.2.1) who state that proponents of the direct causation theory prescribe specific actions for specific emotions.

Zeelenberg et al. The pragmatic "feeling-is-for-doing" approach of Zeelenberg et al. in [215] places emotion as being the primary motivator of goal-directed behaviour. The paper states that whilst the feedback theory of emotion advocated by Baumeister et al. in [11] and [10] may be correct (since it provides a person with information about progress towards a goal) it could also be the case that emotion provides a way to achieve the goal in question. The "feeling-is-for-doing" approach proposes that emotions are elicited when an event or outcome is relevant for a person's particular concerns or preferences. The emotion(s) elicited will then prioritize behaviour that addresses these concerns accordingly. Zeelenberg et al. present details of two experiments that provide evidence for the validity of the approach proposed.

The first experiment by Nelissen et al. [132], details the effects of guilt and fear in context of social dilemmas. According to Zeelenberg et al., [132] contains the best example of goal-activation mechanisms becoming activated when emotional states are induced. The results presented in [132] illustrate how fear and guilt can be used to respectively reduce and increase cooperative behaviour. The study discovered a significant interaction between a participant's emotional state and their social values. Fear decreased the likelihood of pro-social players (players whose behaviour tended to be more altruistic) cooperating whilst guilt increased the likelihood of cooperative behaviour from pro-self players (players whose behaviour tended to be more selfish). In [132] it is stated that fear activates a goal to avoid personal risk but guilt activates a goal to make-up for transgressions. According to the results obtained, goals are only altered if the goal activated by the emotion is not already activated by virtue of the social-values of the player. Therefore, fear only has an effect on pro-social players as pro-self players have a constant, long-term goal to avoid losing to the opponent anyway. Conversely, guilt only has an effect on pro-self players as pro-social players have a constant, long-term goal to keep the opponent's interests in mind thus avoiding transgressions. This would indicate that emotion (in this case, fear) does have an effect on behaviour that is to be performed in the future. It can also be asserted that emotions arise from a consideration of behaviour (as with guilt) and this in turn can have a direct effect on subsequent intentional behaviour. Such a result gives validity to the assertion that emotions both occur before and after intentional behaviour and do in fact directly motivate such behaviour.

The second set of experiments mentioned by Zeelenberg et al. are conducted by Hooge et al. and are documented in [39]. In [39], two forms of shame are defined: endogenous and exogenous shame. Hooge et al. posit that shame motivates pro-social behaviour when it is relevant for the decision at hand (endogenous shame) but that shame has no such effect when it is not relevant for the decision at hand (exogenous shame). The decision to study shame was taken due to its classification as a moral emotion by Haidt [82] who proposes that shame is linked to the interests of others and is responsible for motivating pro-social behaviour[3]. By using three different emotion inductions and two different dependent measures, endogenous shame does indeed appear to motivate pro-social behaviour. After imagining shame with a scenario (experiment 1), pro-self participants displayed increased levels of pro-social behaviour towards the audience in a social dilemma game and these findings were replicated when participants recalled an event involving shame (experiment 2). When experiencing shame after a failure on a performance task (experiment 3), pro-self participants acted pro-socially towards a laboratory audience. Finally, the results from experiment 4 showed that this effect could be generalized beyond social dilemmas to helping tendencies in everyday situations. Ultimately, this experiment allows for the assertion that emotion (in this case endogenous shame), does have a direct effect upon subsequent intentional behaviour

[3]Further links between shame and pro-social behaviour have also been proposed by Emde and Oppenheim in [52] and by Goldberg in [75].

and it is this notion that I will take forward when functionally modelling emotion for computational use.

3.2 Emotion-Rationality Dichotomy

As this thesis proposes the integration of emotion into the decision-making of agents, it is inevitable that any work presented with respect to it is met with questions about the supposed one-dimensional spectrum upon which rational and emotional decision-making lie at extreme ends (this spectrum is illustrated graphically in figure 3.2). Baumeister et al. highlight this issue in [11] where the section entitled "*Emotions and Irrational, Self-Defeating Behavior*" asserts the existence of a stereotype where emotion causes so called "rational" decision-making to be skewed resulting in people engaging in foolish and at times, self-destructive, acts. If this stereotype is correct then it follows that the influence of emotion upon decision-making procedures should be minimised.

The amount of literature dedicated to discussing and attempting to resolve the disparity between emotion and rationality is extensive and includes works by Elster [51], Pham [146] and Pfister and Böhm [145]. The problem in settling this argument is caused by the wide-array of contexts in which these terms may be used. In [213], Wooldridge and Jennings state that, since the behaviour of an agent should be autonomous, it is a tacit assumption that any action performed by such an agent should be the result of a rational decision-making process and therefore the action should be objectively rational. The difficulty in this assumption is defining the term "rational" since the context in which it may be used is so broad. This makes it difficult to reconcile the notions of "rationality" and "emotional" in the context of decision-making since definitions could be altered so that an emotional decision-making process becomes infeasible for use by agents (since it may be considered to be irrational since a rational decision process has not been used).

Therefore, the issue to address is that of *context* i.e. the context in which I will define the term "rationality" (note that I *only* consider the definition of "rationality" in this section as "emotion" will be defined in chapter 5). Consequently, I present and discuss some contexts in which the term "rational" may be used in section 3.2.1 and by using this discussion as a basis, I will then specify the context and definition for "rationality" in this thesis. Following this I will then clarify how the two supposedly opposing concepts of "rationality" and "emotion" can be resolved in section 3.2.2.

3.2.1 Definition of Rationality

A suitable starting point for selecting a definition of rationality in terms of decision-making is the work of Pham, [146], where the existence of three distinct types of rationality are proposed: *material*, *ecological* and *logical*. Sen defines material rationality in [171] as rationality typified by the *homo economicus* model of agency discussed at length in chapter 2. The decision-making of a materially rational agent then centres

around the maximisation of its individual pay-off and consequently, this form of rational decision-making is utilised most notably by classical economic theory. In [146], Pham argues that ecological rationality stands opposed to material rationality in that, an agent who is ecologically rational acts in a way to promote values other than material pay-off (usually a value that benefits others by its promotion). For example: an ecologically rational agent may strive to achieve societal goals or promote moral standards resulting in an ecologically rational agent's decision-making process taking into account a wider range of values other than simple monetary pay-offs.

If rationality is then defined in a material context then an ecologically rational agent's decision-making may be described as "irrational" in that context. If on the other hand an agent's decision-making is appraised in an ecologically rational context then actions such as "taking one for the team" are perfectly rational. An example of this has already been discussed in chapter 2, section 2.4.1.1 where Fehr and Gächter in [56] allow participants in their public goods games to punish free-riders at a cost to themselves. If the rationality of a punisher's decision-making in Fehr and Gächeter's investigation is defined in a materially rational context then their decisions are completely irrational. This is because punishment is costly and does not directly impart an increase of material wealth to the punisher. If the decision-making of punishers is appraised in an ecological context that includes other values such as fairness, considerations of emotional states etc., then such decisions may be considered rational. This is because free-riders evoke anger in other players (as outlined in [56]) resulting in players exacting punishment upon them, maybe as a result of some moral standard being violated. Note that if ecological rationality is ascribed to, punishment for satisfaction and punishment in order to correct behaviour are two completely different rational justifications.

Defining rationality in the context of material and ecological considerations still separates the two contexts whereas I am looking to unite the two together in one definition of rationality. Fortunately, the third type of rationality identified by Pham in [146], logical rationality, allows this goal to be achieved. Kahneman ([97]) defines logically rational agents as agents whose decision-making respects certain logical principles such as transitivity in context of beliefs/judgements/choices/actions. Most pertinent here is the logical principle of *consistency*: if an agent's reasoning is facilitated by Bratman's BDI model [24] then an agent would be said to be irrational if it believes p and desires p but acts intentionally to achieve $\neg p$. Such decision-making is irrational as there exists a contradiction in the logic used: if the agent believes p and desires p then acting intentionally to achieve its negation is inconsistent as p cannot be achieved or maintained in this way. What is important is that the rationality of an agent's decision-making is not defined in the context of specific desires such as wealth or emotion, it is instead defined in the context of logical consistency over any desire an agent may have. Therefore, irrationality and emotion are not synonymous with each other so long as the following two conditions hold:

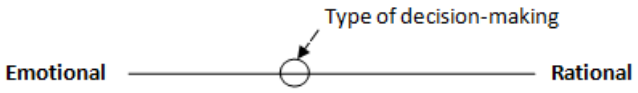

FIGURE 3.2: Graphical representation of the proposed dichotomy between emotional and rational decision-making.

1. An emotion is activated by a stimulus for some given reason that is consistent with the agent's beliefs, desires and intentions (or knowledge base).

2. An emotion motivates an agent to achieve some goal or desire whose achievement is possible and can be achieved by the action performed.

3.2.2 Reconciling Emotion and Rationality

Given the definition of rationality presented in section 3.2.1, it is now entirely possible to consider an agent that uses its emotional state to guide its decision-making as being rational rather than irrational, so long as its decision-making is logically consistent. I agree with Pfeister and Böhm [145] by arguing that, if emotions are appropriate and consistent e.g. an agent, x, defects against an opponent, y, at some cost to x in an iterated Prisoner's Dilemma game due to anger being elicited in x by y, then this is objectively rational if anger is known to cause defection in the context of the game. Likewise, if x cooperates with y at some cost to x in an iterated Prisoner's Dilemma game because gratitude has been elicited in x by y then this is objectively rational if gratitude is known to cause cooperation in context of the game. On the other hand, if x defects against an opponent y in an iterated Prisoner's Dilemma game after having gratitude elicited then x could be objectively described as irrational (if there are no inputs from other emotions that may provoke cooperation). This is because the agent's desires and intentional behaviour is not logically consistent with the agent's beliefs of its current emotional state.

I also argue that the proposed dichotomy between "rational" and "emotional" decision-making in agents (see figure 3.2) should be revised into a two-dimensional space where a combination of the concepts "unemotional", "emotional", "irrational" and "rational" can be used to define an agent's decision-making. It should now be clear that the decision-making of an agent may be described as being a member of four general groups: *emotional-rational*, *emotional-irrational*, *unemotional-rational* and *unemotional-irrational*. Descriptions of these categories are given below along with an accompanying example and graphical representation in figure 3.3 to facilitate explanation and understanding.

- If an agent's decision-making is "emotional-rational" then the agent's decision-making process is driven by a consideration of its current emotional state and the subsequent decision is logically consistent. For example: in an iterated Prisoner's Dilemma game, an agent, x, switches from cooperating with its opponent, y, to defecting against it if this is caused by y defecting against x in the last round and eliciting anger in x. Such a decision is rational as it is reasonable that defection

from y elicits anger in x causing x to subsequently defect. Such decision-making is characteristic of most human players engaged in public goods games (see chapter 2)

- If an agent's decision-making is "emotional-irrational" then the agent's decision-making process is driven by a consideration of its current emotional state but the subsequent decision is logically inconsistent. "Stockholm Syndrome" [78] is an example of such decision-making: an individual that is held prisoner by another who causes them physical or psychological damage can develop positive emotions such as pity, empathy and love towards the captor. Such feelings and their associated intentional actions are irrational as it is not a normal consequence that pain exacted by an individual towards another elicits positive emotions in the victim. Note that the intentional behaviour and the physiological symptoms still exist allowing the emotion to be attributed but these behaviours and symptoms are inappropriate to the stimulus, which would normally elicit some other emotion.

- If an agent's decision-making is "unemotional-rational" then the agent's decision-making process is driven less by the agent's current emotional state and by something else (monetary gain for example). Furthermore, the subsequent decision is logically consistent. For example: in an iterated Prisoner's Dilemma game, an agent, x, switches from cooperating with its opponent, y, to defecting against it. This is because y has defected against x in the last round and x wants to maximise its individual pay-off in this round (and avoid the sucker's pay-off). If y defects again in the next round and x cooperates then x will receive the sucker's pay-off. However, if y cooperates in the next round and x defects then x will earn the highest individual score possible. Such decision-making is the hallmark of game-theoretic utility maximisation.

- If an agent's decision-making is "unemotional-irrational" then the agent's decision-making process is driven less by the agent's current emotional state and by something else (monetary gain for example). Furthermore, the subsequent decision is logically inconsistent. For example: in an iterated Prisoner's Dilemma game, an agent, x, wishes to distribute the sucker's pay-off to maximise its individual pay-off and in the last round its opponent y cooperated. If x then cooperates in the next round its behaviour is irrational as cooperation can never produce the most preferred individual outcome for an agent. Such decision-making may be defined as random as it is not motivated by either emotion or rationality.

3.3 Chapter Summary

As stated in the introduction to this chapter, I intended to address a number of important philosophical questions regarding emotion that had a significant bearing upon the entire

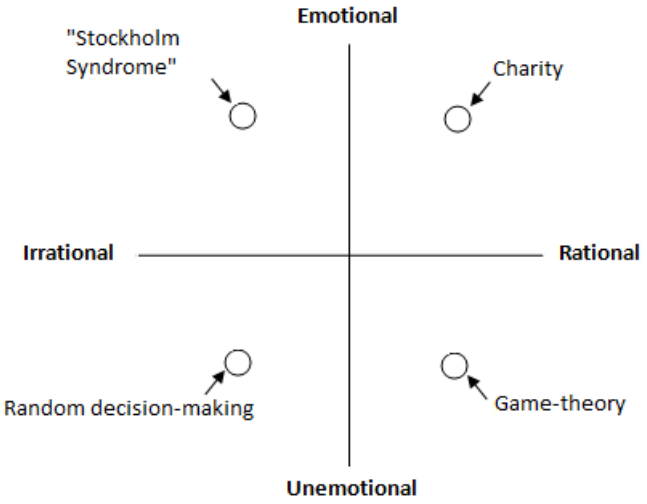

FIGURE 3.3: Graphical representation of the proposed two-dimensional emotional/unemotional and rational/irrational plane with examples of each type of decision-making identified in each quadrant.

thesis I am proposing. The general purpose of section 3.1 was to make my position explicit with respect to the relationship between emotion and intentional behaviour. In section 3.1.1 I presented philosophical accounts of emotion concerned with whether unintentional behaviour is a necessary part of an emotion and whether there is a link between emotion and intentional behaviour. As demonstrated in table 3.1 of section 3.1.1.4, an emotion can exist in the absence of unintentional behaviour so long as a perception and an intentional behaviour are present; in this case an emotion can be said to be simulated.

In section 3.1.2, I sought to make clear my stance concerning when emotions occur in respect to the occurrence of intentional behaviour. Since I am interested in the effects of emotion upon intentional behaviour in social dilemma games, I will take the stance that emotion occurs before intentional behaviour and thus motivates it. Ample support for this view has been given in section 3.1.2 and to argue against theories of emotion which assert that emotion can only occur after intentional behaviour.

The supposed dichotomy between the concepts of "emotion" and "rationality" is addressed in section 3.2 since this is still an issue of contention between researchers[4]. The section provides both a discussion and a visual representation of how I consider emotion and rationality to be related: that instead of being mutually exclusive they are independent of one another. Consequently, the decision-making of an agent may, at any time, be described as being a member of one of four general groups: *emotional-rational*, *emotional-irrational*, *unemotional-rational* and *unemotional-irrational*.

In summary, the salient points to bear in mind before progressing further with the thesis are as follows:

[4]See Scherer's discussion of the topic [168]

- Unintentional behaviour is not a *necessary* part of an emotional experience and an emotion can still be said to exist even without the expression of *unintentional* behaviour. However, without the possibility of *unintentional* behaviour, the best I can achieve with respect to an emotional model of agency is to *simulate* an emotion.

- There is a *functional* relationship between an emotion and its associated intentional behaviour i.e. the elicitation of an emotion can cause an agent to perform some form of intentional behaviour to achieve/maintain some goal.

- I side with the argument that emotions occur *before* and influence the emergence of intentional behaviour. Whilst some emotions may occur after intentional behaviour I will not consider this aspect of emotion and intentionality here.

- The supposed dichotomy between *emotion* and *rationality* does not exist; if decision-making is *logically consistent* then decision-making is rational and so an agent's decision-making can be influenced by emotion and still be logically consistent.

Chapter 4

Modelling Emotion

I argued in chapter 2 that the behavioural predictions of *homo economicus* were insufficient to explain the observed behaviour of *homo sapiens* in public goods games. I also argued that this in-correctness was primarily caused by the selfish, narrow-minded considerations of *homo economicus* with regards to economic utility. However, in almost every case studied in behavioural economics, utility cannot just be narrowly defined as wealth and agents are not as selfish as *homo economicus* predicts. In some cases, considerations of the player's emotions as well as those of others are included in the utility to be maximised. It may well be that this is the integral element missing from the *homo economicus* model of agency that prevents it from correctly predicting the behaviour of *homo sapiens* in public goods games.

The study of emotion in a computational context is commonly referred to as "affective computing" and the topic is extensively covered by Picard in her textbook of the same name [147]. Picard's book largely concerns itself with the investigation of how computer systems may recognise the existence of an emotion in a human user and use that information in some way (change the current user-interface, re-designate CPU time etc.). It is at this point that I will make the division between my work and works such as Picard's clear: in no way at all is this thesis interested in sensing or recognising emotions in human users. As was stated in section 1.1 of chapter 1, the general goal of this thesis is to determine whether or not emotions can be modelled functionally in a computational context. By doing this I aim to analyse and understand the impact of emotions upon the social interactions of agents engaged in public goods games; this will go some way towards addressing the identified shortcomings of the *homo economicus* model. To achieve these goals I begin in section 4.1 by providing a survey of various psychological models of emotion that are used as basis from which computational models of emotion are developed. The main purpose of this section is to investigate how emotion elicitation should be modelled since, in order for an emotion to have an effect on an agent's intentional behaviour, its occurrence must first be provoked.

Moving onwards, this chapter then provides a much more focused discussion of literature pertaining to the computational modelling of emotion and how emotions have been utilised in MAS. There are two very broad categories that such literature may fall

into: logical formalisms of psychological models of emotion and implemented emotional systems. Literature concerning these categories is discussed respectively in sections 4.2 and 4.3. The general purpose of these two sections is to provide a basis for contextualisation of my own work amongst the extensive array of literature concerning computation and emotion.

More specifically, by discussing logical formalisms of psychological emotion models in section 4.2 I intend to explore how an emotional experience, from its elicitation to its end, is modelled using logic. Such models provide clear, unambiguous details of how an emotion is elicited computationally, how an emotion translates into the emergence of an intentional behaviour, how long the emotion lasts for etc. In this section I will also exemplify noteworthy features of these logical formalisms that will be adopted for use in my own research.

The problem with logical formalisms however, is that they are usually rather abstract and so, in section 4.3, I shift my attention towards a discussion of computational emotion models that are physically implemented in computational systems. Whilst this section again elucidates how emotional experiences are modelled and implemented in computer systems, such details are much more concrete allowing for any noteworthy features to be more readily adopted for use in this thesis. In addition, implementation of abstract theories usually requires some simplifications to be made and other design choices to be made; such features should be noted. A pertinent question that also arises from such a discussion is how implemented computational models of emotion are used by computer scientists to further their own research agendas. Therefore, in section 4.3 I will also provide answers this question and by doing so I will provide a basis to consider my own research agenda and pinpoint its novel aspects.

4.1 Psychological Models of Emotion

To this point, I have discussed the explanatory power of emotion in the context of non-rational human behaviour and theories that propose how emotion is capable of producing such behaviour (see chapter 2). In this section I will consider the major theoretical models concerning emotion elicitation and highlight salient aspects of these that will be carried forward into my own emotional model.

The structure of this section takes inspiration from work by Marsella et al. [121] i.e. emotional models are divided into three types or groups: appraisal models (section 4.1.1), dimensional models (section 4.1.2) and anatomical models (section 4.1.3).

4.1.1 Appraisal Models

Appraisal models of emotion are perhaps the most widely used in the fields of psychology and computer science. The core strength of such emotional models is in the subjective basis of their mechanism. For example: whilst the sight of a cat may elicit intense distress in someone who has been attacked by one, the same cat may elicit intense joy

in someone who has had positive experiences with cats. In this way, the cat is appraised against some subjective criteria resulting in the same object potentially eliciting very different emotional responses for different people. It should be noted that the example just provided is, of course, an oversimplification since there are usually more criteria other than past experience considered when appraising a given situation in an emotional context but it serves to highlight the main benefit imparted by appraisal models of emotion.

4.1.1.1 The Ortony, Clore and Collins Model

The OCC model of emotion [142] (named after its creators: Ortony, Clore and Collins), is perhaps the most applied psychological model of emotion with respect to the modelling of emotion in computer science. Its adoption by computer scientists as the *de facto* method of modelling emotion is largely due to its tractability and use of concepts that are closely linked to those used by the agent systems community.

The global emotional structure of the OCC model proposes twenty-two emotions, with an emotion defined as a valenced reaction of an agent to the consequences of an event (relevant to the agent itself or another), the actions of an agent (either the agent itself or another) or some aspect of an object. As stated, the flexibility of the model's structure is impressive as it takes into account the subjectivity of an agent when appraising an event, agent or object. For example: Ortony at al. state that, whilst a car may be treated as an object, it may also be treated as an agent since an agent (according to the model), may be anything from a person to an inanimate object. What matters is how the object is construed by a person: a car may be viewed as an agent e.g. if a person's car breaks down they may blame the car in much the same way as they would another human being (cf. Dennett's intentional stance [41]). Therefore, a person would disapprove of a car if they construe it as an agent whereas they would dislike the car if it were construed as an object. The OCC model's global structure is presented in figure 4.1; the valenced reactions to each aspect of the environment detailed above can potentially give rise to a broad range of affective reactions. As noted by the authors, the primary determinant of whether or not these affective reactions are experienced as emotions is their *potential* value and as such, the model goes into great detail regarding variables that affect an emotion's potential. The idea that the potential of an emotion must reach a certain threshold before it affects an agent's intentional behaviour, is an especially important feature of the functional correlate of emotion outlined in this thesis.

Ortony at al. posit the existence of an implicit or virtual appraisal structure composed of a person's goals and the links connecting these goals. The structure is considered to be dynamic (goals and the links between them are added/deleted/modified frequently, useful for agents inhabiting dynamic environments) and the layout is proposed to be that of a lattice rather than a tree. This notion of goals is part of the reason why the model is adapted for use so readily by computer scientists, especially those working in the field of agent/MAS. The OCC model recognises the work by Schank and Abelson [166]

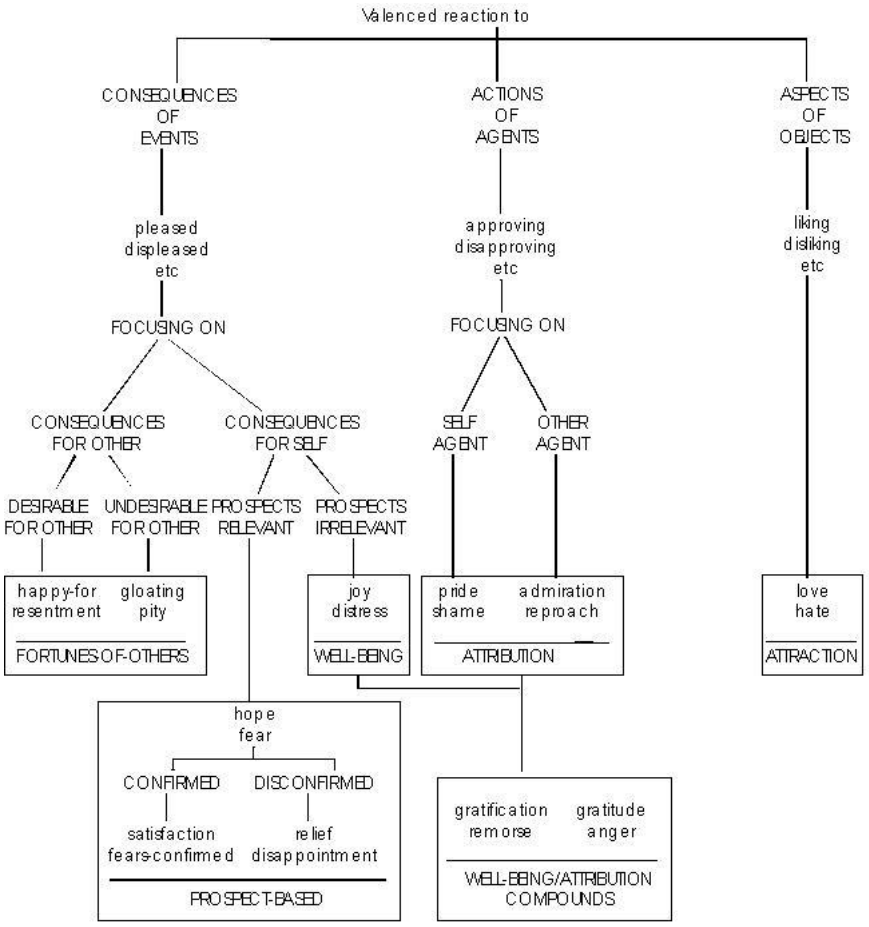

FIGURE 4.1: Graphical representation of the global structure of emotion according to the OCC model [142].

which broadly divides goals into achievement, satisfaction and preservation goals among others. Of these goals, the OCC model makes use of three, namely:

- Active-pursuit goals - represents goals that people want to achieve themselves by putting in some effort. For example: "climb Mount Everest", "apply for job" etc. Subsumes Schank and Abelson's achievement/entertainment/instrumental/crisis goals.

- Interest goals - represents goals that people want to see happen since it benefits their lives in some way. For example: "preserve own health", "see favourite sports team win" etc. Subsumes Schank and Abelson's preservation goals.

- Replenishment goals - represents goals that become increasingly urgent until they are achieved. After achievement, the urgency of achieving these goals drops to zero. For example: "fill car with petrol", "eat food" etc. Subsumes Schank and Abelson's satisfaction goals.

There is a noticeable parallel here between the notions of achievement/maintenance goals in MAS and active-pursuit/replenishment goals in the OCC model; such parallels facilitate the use of the model in agent systems. The OCC model stipulates that consequences of events should be evaluated by agents in the context of their goals, for example: I may not be able to start my car in order to attend a job interview and due to this, I may fail to achieve the goal of landing the job. By appraising the event "car has broken down" in the context of my goal to "land the job" I may then kick the car whereas if the event "car broken down" was not appraised in the context of my "land the job" goal, I may just walk away. It should also be mentioned that this appraisal process allows for the elicitation of opposing emotions towards the same event/action/object but only if the goal context in which that event/action/object is appraised in differs. For example: I may dislike water if my goal is to avoid getting wet but I may like water if my goal is to quench my thirst.

The appraisal structure proposed by the OCC model also necessitates the inclusion of standards and attitudes along with goals. It would appear that standards represent the values of the appraiser, for example: "people should not judge others without meeting them", "people should not lie to others" etc. Attitudes are intended to represent the tastes of a person, for example: "I do not like pineapples", "I like the colour green" etc. Ortony et al. assert that attitudes differ from standards in that they do not require justifications for their existence i.e. if I like the colour green that is all that needs to be said on the matter. This facilitates the interpersonal variation that any theory of emotion must address. The inclusion of standards and attitudes is deemed necessary by Ortony et al. since standards form the basis upon which we evaluate the actions of others and attitudes form the basis of evaluating aspects of objects.

One of the key ideas in the OCC model is that emotions are not experienced until a certain *threshold of potential* has been reached. The mechanism proposed by Ortony et al. that affects emotion potential is best described by the authors themselves:

> "In the general case, the kind of intensity mechanism that we envision is something like the following: The eliciting conditions for some emotion are satisfied, resulting in the *potential* for that emotion to be experienced. We assume that there is a context-sensitive emotion-specific threshold associated with each emotion so that the emotion will only be experienced if its threshold is exceeded."

Ortony et al. then go on to state that the intensity of the emotion experienced is dictated by how much the emotion's associated potential threshold has been surpassed; the magnitude at which the emotion's potential threshold is surpassed is influenced by various intensity variables. This distinction between *potential* and *intensity* may be confusing so I will make use of an example to clarify. Let us consider the emotion *joy* as defined in [142]; the equation 4.1 describes the activation rule for this emotion.

$$IF(desire(p,e,t) > 0)\{$$
$$joyPotenial(p,e,t) = f_j(desire(p,e,t) + globalIntensity(p,e,t)) \quad (4.1)$$
$$\}$$

The "if" statement in equation 4.1 states what conditions must exist in order for emotions of type joy to exist (the desirability of some event, e, must be positive for a person, p, at time, t). If this condition is met then the $joyPotential$ variable (the emotion's *potential*) in context of the event e for person p at time t is set to the result of a function used specifically for joy emotions (f_j) that sums together the desirability of the event e for the person p at time t and the weights/values of the global intensity variables specific to the event e for person p at time t. If the value of $joyPotential$ is then greater than the threshold for emotions of type joy then joy is elicited in the agent. The *intensity* of the emotion is then equal to the difference between the joy's current potential and threshold. This intensity value may be then used to ascribe a lexical token to the emotion experienced i.e. "happy" for joy whose potential is only slightly greater than its threshold (less intense joy) and "ecstatic" for joy whose potential is much greater than its threshold (very intense joy). As stated, this concept of *thresholds* is extremely important with respect to the thesis and therefore the details of this should be kept in mind.

4.1.1.2 Frijda's Model of Emotion

Frijda's magnum opus [67], outlines an emotional model that is predominantly functional and slightly more complex than the model proposed by Ortony et al. (see section 4.1.1.1) and this model is further refined in a joint paper authored by Frijda and Swagerman [68]. Emotion elicitation in Frijda's model is orientated around an individual's satisfaction conditions (defined as particular internal/external states) and concerns (motivational constructs that serve to motivate an agent to achieve/avoid/maintain their satisfaction conditions) of which there are two sub-classes: source and surface concerns. Source concerns are satisfaction conditions that pertain to states of the individual e.g. security, hunger, individual finances etc. Surface concerns on the other hand are satisfaction conditions that involve some external object e.g. love for another person, despair at human rights violations, fear of spiders etc. These two notions have parallels in the OCC model [142] but the OCC model's definitions appear to be more specific since source concerns are analogous to the definition of goals proposed by Ortony et al. and surface concerns could be interpreted as either standards or attitudes according to the OCC model. Frijda avoids the use of terms such as "goal" and "motive" as they imply an awareness of a future state and/or some form of activity which is sometimes inappropriate with respect to the conditions under which some emotions arise.

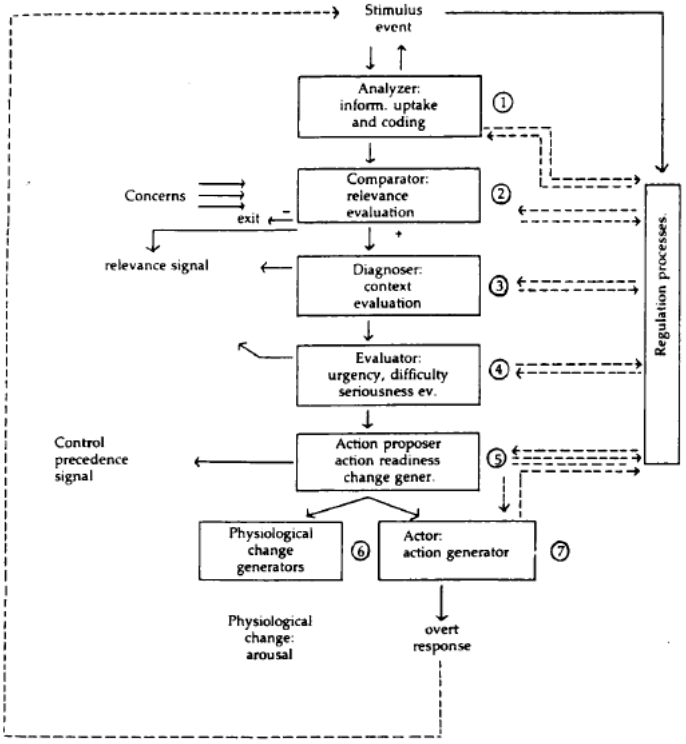

FIGURE 4.2: Frijda's model of emotion [67].

In [68], Frijda and Swagerman propose that emotions are part of a larger system of concern realisation and argue that such a system is necessary since human beings and some other animals (no particular animals are mentioned) need to realise multiple concerns with limited resources in the context of dynamic environments[1] (the flow of control throughout this system is shown in figure 4.2). In much the same way as Ortony et al., Frijda models emotions so that their effect is to generate *action tendencies*. According to Frijda, programs of applicable actions may be performed by the actor to deal with the current situation in order to achieve/avoid/maintain a satisfaction condition. If, however, the emotion's intensity is not strong enough (or does not surpass its associated *threshold*, to use OCC model terminology), no action is taken. Unfortunately, the intensity calculation used by Frijda is not detailed (to the best of my knowledge) in either [67] or [68] although it is stated that, the more important the concern, the greater the emotion intensity elicited if that concern's achievement/maintenance is facilitated/thwarted.

Finally, Frijda supports the idea that opposing emotions may not be felt towards the same concern since a concern either triggers a positive or negative reaction, never both. However, opposing emotions may be felt as long as the relevant concern is different; this idea is analogous to that proposed by Ortony et al. in [142].

[1]This proposition is analogous to Simon's concept of "bounded rationality" [174].

4.1.1.3 Lazarus' Model of Emotion

According to Lazarus [108], emotion elicitation originates from the appraisal of significant events that effect personal goals and well-being and the work presented is a rather fine-grained analysis of mechanisms that allow for appraisals to be made with respect to emotions. Aside from this, Lazarus takes the exceptionally strong position that emotion is always a result of cognition i.e. cognition is both a necessary and significant condition for emotion. What is characteristic about this viewpoint is, as Lazarus himself puts it:

> "Many writers who accept comfortably the idea that cognition is sufficient reject that it is necessary."

"Sufficient" in this particular context means that thoughts are capable of producing emotions and "necessary" entails that emotions are not capable of being elicited without some degree of thought. Furthermore, Lazarus makes clear his functional standpoint upon the relationship between cognition and emotion i.e. the relationship is bi-directional, cognition generates an emotion and this emotion in turn may generate cognition.

The distinction made between appraisal and knowledge by Lazarus is important in understanding the proposed mechanism for how emotions are elicited. According to Lazarus, "knowledge" is defined as what a person believes about how the world works in both a general and specific context whereas "appraisal" is defined as the use of significant knowledge about a given situation to evaluate how that situation may affect personal well-being. Thus, this appraisal of knowledge in context of personal well-being gives rise to a recognition that well-being may be affected (either positively or negatively) which in turn gives rise to an emotion.

There is, unfortunately, no mention of emotion thresholds or discussions of emotional intensity in Lazarus' emotional model. Instead, the paper goes on to discuss different types of cognition and how they may/may not play a part in emotion elicitation. Such discussions are of no concern here as this is not a psychological thesis and so no discussion of such topics will be offered here.

4.1.1.4 Scherer's Model of Emotion

The work produced by Scherer in [167] contains an in-depth consideration of the appraisal process associated with emotion elicitation. Scherer proposes a number of stimulus evaluation checks or SECs i.e. a set of criteria that underlie the assessment of how significant a stimulus event is for an organism. These SEC's are organised in terms of four appraisal objectives which concern the major types/classes of information that an organism requires in order to prepare an appropriate reaction to some object/event. These four appraisal objectives are outlined below:

- Relevance detection: appraises events and how they may affect the self.

- Implication assessment: appraises stimuli in context of whether or not it furthers/endangers the appraiser's survival/adaptation in/to the environment.
- Coping potential determination: determines what responses are available to the appraiser and what consequences would be entailed. Produces an estimated degree of coping potential for the most beneficial response available in the current context.
- Normative significance evaluation: if the appraiser is of a social species this evaluation is responsible for determining how the other members of society would evaluate the actions taken in response to an emotionally-stimulating event.

The appraisal process detailed here is much more specific with regards to the appraisal components than any of the other emotion models already discussed. Indeed, such an appraisal mechanism may be too fine-grained and detailed to be considered practically useful in order to create a functional mechanism that mirrors emotion. This is especially true if Scherer's idea that such an appraisal is not made once is considered: instead, subsequent appraisals will apply as the environment is dynamic and thus the action taken in response to an event at one stage in time may not be the best response to the same situation at some later point in time. It is however notable that a large amount of empirical evidence supports Scherer's claims that emotional appraisal does utilise some mechanism akin to the SECs proposed above (if not SECs themselves). The model is interesting in that it is the first one discussed that explicitly takes into consideration the social consequences of emotionally driven behaviour.

It would seem also that these SECs may be used in order to determine an emotion's intensity if it is elicited but as with Frijda and Lazarus' emotion models discussed in sections 4.1.1.2 and 4.1.1.3 respectively, there is no specific equation or mechanism given. Scherer does however entertain the notion of emotional *thresholds* and posits that an emotion's intensity must reach a certain threshold before its effects are manifest (the fact that emotions serve to directly influence intentional behaviour would indicate that Scherer takes a functional view of emotion). Also, in keeping with Frijda's and Ortony et al. (see section 4.1.1.1), is Scherer's hypothesis that the four appraisal objectives above are given different weights in the appraisal process depending upon the context of the current situation. It is proposed that the output of emotion are *action potentials* (in keeping with Frijda's emotional model) so, whilst an individual may be consciously experiencing some emotion, an action is not guaranteed to be performed as a result.

4.1.2 Dimensional Models

Dimensional models of emotion are characterised by the identification of various scales or dimensions that determine an emotional state depending upon where the emotion is located on the dimensions proposed. For example: Schlosberg in [169] proposes two dimensions: unpleasantness-pleasantness and rejection-attention. In later work, [170], Schlosberg proposes a third dimension: sleep-tension, that describes the activation level

or intensity of an emotion. In [1], Abelson and Sermat investigate the applicability of the "rejection-attention" scale and found it to be redundant since the "sleep-tension" scale is much better at predicting dissimilarity data. Research by Engen, Levy and Schlosberg [55] on the other hand confirms the validity of the three-dimensional model by comparing the judgements made by students and newspaper readers upon emotionally charged pictures on the three-dimensional scales proposed by Schlosberg.

The drawback of dimensional models is that the origin of emotion elicitation is not made explicit as it is for appraisal models of emotion (see section 4.1.1). So, whilst it may be asserted that the pleasantness of a stimulus has a bearing upon the emotion produced, the stimulus itself is not specified and neither is the appraisal process that determines how pleasant the stimulus is. One advantage of dimensional models however, is that calculations pertaining to emotional intensity are included although there is little consensus as to what these intensity variables should be (as is also the case with appraisal models of emotion). Furthermore, there is no notion of *thresholds*, an integral factor in most appraisal models of emotion.

To summarise, dimensional models may be particularly useful for modelling emotion elicitation in the context of general stimuli and investigating the ubiquity of emotional appraisal amongst human beings. Unfortunately, the drawback of dimensional models is twofold; they are not so useful at prescribing how to model emotion elicitation proper and there is a worrying lack of consensus as to what dimensions need to be considered. As a consequence, dimensional models will not be considered as a viable model upon which to build the mechanism I am proposing.

4.1.3 Anatomical Models

Unlike the more abstract appraisal and dimensional models of emotion considered in sections 4.1.1 and 4.1.2 respectively, anatomical models of emotion place a great emphasis upon the anatomy of the body, neurobiology in particular, in their explanations of emotion. Damasio's emotional model [34, 35] is perhaps the most notable anatomical model that exists and has already been noted in chapter 1 although it was not made clear that this model was anatomical in nature. In addition to Damasio, LeDoux is also prominent in the field of anatomical models of emotion and outlines his own model in [110].

Therefore, anatomical models of emotion are more concerned with the hardware of the brain which facilitates the operation of these abstractions rather than the abstractions themselves. Consequently, I will not provide any further discussion of anatomical models of emotion as they offer little in the way of computationally applicable insight (from the point of view of this thesis) into how emotions are elicited.

4.2 Logical Formalisms of Emotional Models

In this section I intend to provide a relatively brief discussion of the most prominent logical formalisms of psychological emotion models. With this in mind, the discussion presented in this section will only consider literature whose aim is to model emotion elicitation and the consequences of experiencing emotion computationally. The logical formalisms that are taken into account in this section are based upon appraisal models of emotion that were outlined in section 4.1.1 of this chapter. Indeed, the logical formalisms presented here exclusively formalise Ortony et al.'s model of emotion [142] that was discussed in section 4.1.1.1. This is not due to selective filtering on my part: rather, most logical formalisms and implemented computational models make use of the OCC model due to its tractability, degree of abstraction and ease of translation into computational terms. Therefore, it would seem appropriate that any model of emotion proposed by myself should also take inspiration from this psychological model with respect to modelling the emergence and effects of emotions.

4.2.1 Steunebrink et al.

Perhaps the most notable researchers with respect to the field of computational instantiations of emotional models are Stenuebrink et al. whose work on providing logical formalisms of the OCC model of emotion spans multiple papers: [36], [186], [187], [189], [188] and [190]. In this section however I only focus upon [187] since it it quite broad in its treatment of emotional modelling and is most representative of their work. The paper takes into account: the modelling of emotion eliciting conditions, the effect of eliciting an emotion, the modelling of emotion intensity and the experience associated with an elicited emotion.

The core intention of Steunebrink et al. in this paper (and indeed in the others mentioned) is to model and formalise human emotions in computational logic. The authors state that their goal is not to develop a specific computational model of emotions, but rather to develop a logic for studying emotional models and they begin with the OCC model. Stenunebrink et al. assert that the OCC model is suitable for such logical formalisation because of its concise hierarchy of emotions and detailed specification of conditions that elicit each emotion in terms of the emotion's focus (recall that there are three emotional foci: events, actions and objects, see section 4.1.1.1). The concepts of events, actions and objects correspond to notions commonly found in the context of agent models and are translated respectively into goal accomplishment/failure, plans (consisting of domain and sequential compositions of actions) and agents. Accordingly, this makes the OCC model suitable for use both in the deliberation and practical reasoning of artificial agents. It is important to highlight this point as I am concerned with how emotion affects the intentional behaviour of agents functionally and, according to Steunebrink et al. the OCC model is a well-suited candidate for modelling such aspects.

One of the most important aspects of this logical formalisation is that an emotional triggering fluent or $EmoTrig$ is defined for each of the twenty-two emotions of the OCC model. Emotions are modelled as opposing pairs or as positive and negative antipodes (hope and fear are modelled as an opposing pair, for example). It is important to note that Steunebrink et al. have formalised the emotions defined by the OCC model to enable agents to experience opposing emotions simultaneously but only if those opposing emotions have a different foci. For example: I may feel joy since I am to receive a cash windfall but I may also feel distress about a work-related issue simultaneously. The foci of the two emotions in the example presented are different: I cannot feel simultaneous joy and distress about the cash windfall event and neither can I feel both joy and distress about the work-related issue. This approach to simultaneous opposing emotion experience differs from the OCC model which allows opposing emotions to be experienced simultaneously with respect to the same event/action/object as long as the appraisal is performed in context of a different goal (see section 4.1.1.1). This point is worth noting since I will need to make clear what my approach to dealing with simultaneous opposing emotions is since it is integral to the correct functioning of any emotional model.

In order to understand the formalisms presented by Steunebrink et al., some of the notation used should be clarified:

- i - denotes the agent being considered.
- j - denotes another agent distinct from i.
- α - denotes an action performed by an agent. The action is performed by the agent that precedes the occurrence of α in the logical formalism. So, in the following formalism, α is the action performed by agent i: i, α.
- κ - denotes a (sub)goal of an agent. The agent that the goal belongs to is given by the identifier that precedes the occurrence of κ in the logical formalism. So, in the following formalism, κ is the goal of agent i: i, κ.

Steunebrink et al. use gratification as an example of how emotions are translated from their OCC model definition into the corresponding logical formalisation. The eliciting condition of gratitude as defined by the OCC model is as follows: "*gratification is approving of one's own praiseworthy action and being pleased about the related desirable event*". According to [187], gratification is defined with respect to an agent i and takes into account both an action α and the accomplishment of a goal of the agent κ. Gratification can therefore be logically modelled as: $gratification_i(\alpha, \kappa)$ and is read as: agent i has performed action α accomplishing (sub)goal(s) κ eliciting gratification. It should be highlighted at this point that goals (κ) and actions (α) are not explicitly defined by the implementer as being undesirable/blameworthy or desirable/praiseworthy rather, the agent itself makes these judgements. Therefore, whilst a goal such as "become a millionaire" may be desirable for one agent, it may be completely undesirable for another. Continuing on from this point, a desirable event is translated as the

accomplishment of one or more sub-goals by an action whereas an undesirable event is defined as being the opposite i.e. the non-achievement of one or more sub-goals by an action. The praise/blameworthiness of actions are also translated in a similar way: for an agent i to deem an action, α, as being praiseworthy, it needs to be determined if α has accomplished i's desirable goal, κ. If an action, α, has indeed accomplished a desirable goal, κ, then α is deemed as being praiseworthy.

I shift focus now onto a consideration of the quantitative aspects of emotions and how they are are addressed in Steunebrink et al.'s logical formalism. Generally speaking, these emotional aspects are defined in terms faithful to the OCC model: potentials, thresholds, and intensities. Therefore, the potential of an emotion can be thought of as how excited an emotion currently is, the threshold of an emotion denotes what an emotion's potential must equal before that emotion is said to have been activated/elicited and intensity is the value of the emotion's threshold subtracted from its current potential. So, if an emotion such as anger has a current potential of 6 and its threshold is 4 then the current intensity of anger is equal to 2. If the threshold of an emotion is greater than its current potential then the intensity of that emotion is equal to 0 (the minimum value of an emotion's intensity must be 0 and cannot be negative).

In total, 22 emotional experience fluents (*EmoExp*) are defined by Steunebrink et al.; one for each of the 22 emotions modelled. An *EmoExp* is satisfied if and only if, the intensity associated with it is greater than zero at the current time. Therefore, if the potential of a newly triggered emotion is less than its threshold then the *EmoExp* is never satisfied and the effects of the emotion are not experienced. As stated in sections 4.1.1.1, 4.1.1.2, 4.1.1.4: these notions of emotional thresholds/intensities and the relationship between them are extremely important as this relationship facilitates the computation of emotion by an agent system and special attention should be given to its usage.

With respect to the effects of emotion, it is best to turn attention towards [186] as there is no clear discussion of this issue in [187]. In [186], Steunebrink et al. use the emotions of hope and fear as examples to describe the effects of an emotion being activated with respect to an agent. Hope and fear are defined as "complementary" emotions by the authors meaning that their intensities must sum to a constant if their foci are the same. Consequently, specifications of these emotions must consider both hope and fear so Steunebrink et al. outline two possible combinations of these emotions.

The first deals with a situation where an agent experiences hope but not fear. In this case, the OCC states that when an agent is hopeful with respect to a plan, π, and a goal, κ, it is pleased about how the plan is progressing. So, if an agent experiences hope without fear then the agent should keep its current intention and commitments (its plan, π) with respect to its goal, κ. Heuristically speaking, the agent should not deliberate about alternative plans (π') to achieve its goal, κ, when it is hopeful.

The second combination of these emotions entails that both hope and fear are experienced simultaneously with respect to a plan, π, and a goal (since the emotions are

complementary they are both allowed to exist simultaneously with respect to the same foci). According to the OCC model, an agent experiences fear when it is displeased about the prospect of an undesirable event i.e. the non-achievement/achievement of a desirable/undesirable goal. In this case, the agent should produce a new plan, π', in order to achieve/avoid achievement of the desirable/undesirable goal but only when the intensity of fear is greater than the intensity of hope. This, however, is a specific example and it would appear to be down to the discretion of the implementer to determine what the effect of an emotion should be. This is important as it gives a certain degree of subjective freedom with respect to the consequences of emotion elicitation (an important issue when modelling emotion).

The logical formalism developed by Steunebrink et al. prescribes that the consequences of an emotion should have an effect on the emotion's focus, or if there are more than one, on one of the emotion's foci (if the emotion's current intensity is greater than or equal to its associated threshold). Generally, Steunebrink et al.'s work on this subject shows how emotions can be used to resolve what might be a complex decision: in the case of hope and fear, whether to continue with a current plan or produce another.

The issue of intensity decay is tackled in some detail by Steunebrink et al. in [187]. To calculate the decay of an emotion's intensity the authors propose that a reasonable default choice is an inverse sigmoid function which depends on several parameters being set so as to give the function the shape desired for a particular emotion. The use of the term "desired" here makes it clear that there is no attempt by the authors to provide a definitive emotional decay function for all emotions specified. The sigmoid function mentioned is a function of time (x) and is outlined in equation 4.2 below. A description of the parameters also follows:

$$int(q_i, t_0, \mu, \delta)(x) = \frac{q_i}{1 + (e^{x - t_0 - \mu)\delta}} - c \tag{4.2}$$

- q_i: the initial intensity of the emotion
- t_0: the time at which the emotion was initially triggered
- μ: the half-life time of the emotion's intensity
- δ: the fall-off speed of the emotion's intensity
- c: used to cut off the intensity for a large enough x

Functions for calculating the half-life (μ) and fall-off speed (δ) of an emotion are not provided by Steunebrink et al. although they do suggest that fall-off speed should be set to 1 to create a "normal" sigmoid curve. Unfortunately though this number appears to be purely arbitrary and no psychological evidence or otherwise is provided as a basis for choosing this value. There is also a distinct absence of emotional potential/threshold calculations defined in the paper. Consequently, it would appear that such calculations are application-dependent i.e. such calculations are open to interpretation by developers

and the best method of calculation may differ depending upon the model's application domain.

4.2.2 Adam et al.

In a similar vein to Stunebrink et al. in section 4.2.1, Adam et al. also attempt to provide a logical formalism of the OCC model in [2]. The author's reasons to model the OCC are numerous, the most salient being that the OCC model is widely used in the design of emotional agents. This is due to its simplicity and ease of implementation which matches the expectations and needs of computer scientists. It would therefore seem that the combination of the OCC model's finite set of appraisal variables suffices for current applications. The authors also agree with the proposition of the OCC model that any emotion must be consistently valenced since this confers the advantage of providing a clear test to differentiate emotions from closely related notions that are not valenced. Furthermore, the authors argue that valence is something naturally captured by logic and computer science resulting in the OCC model being well adapted particularly for a logical formalisation. In addition, Adam et al. notes that the OCC model has a simple and elegant tree structure which facilitates tractability[2]. The OCC model also uses concepts that are well understood in logic such as beliefs, desires and standards thus making the formalisation task easier. Finally, the authors argue that the OCC model is quite exhaustive, an important feature if one is to design robust and versatile agents i.e. agents that can emotionally react to a broad variety of stimulus.

The aim of [2] is to account for relationships between the different components of emotions and relationships between an agent's emotions and their performed actions (these topics are not accounted for in the OCC model). As such, Adam et al.'s formalism attempts to describe an agent's mental attitudes such as beliefs, goals and desires with the aid of the *EL* logic and the BDI-focused framework developed by Herzig and Longin in [88].

Adam et al.'s formalism is restricted to the modelling of emotion triggering conditions thus, they devote no attention to the logical formalism of emotional experience (this is covered by Steunebrink et al., see section 4.2.1). The authors also restrict their formalism to event and action based emotions (objects are not considered) and consequently, the salient variables modelled are the desirability/undesirability of an event and the praise/blameworthiness of an agent's action. It is argued that the desirability of an event is closely tied to the concept of utility i.e. the more positive the utility conferred by an event, the more desirable the event is and that the same event may be desirable/undesirable depending upon the circumstance i.e. the event is valenced. The authors recognise that this is problematic as such a feature makes a logical formalisation difficult due to its necessitation of "*either a paraconsistent notion of desirability such*

[2]This argument is made despite Ortony et al.'s assertion that the OCC model is more of a lattice than a tree (see section 4.1.1.1).

that 'φ is desirable' and '$\neg\varphi$ is desirable' are consistent, or a binary notion of desirability that is relativized to goals".

Adam et al. tackle this issue by shifting the focus of an agent's appraisal of an event or action to the consequence of that event or action in terms of the agent's current goals. So, whilst an agent that has lost its job may have distress elicited due to the goal of "being employed" not being achieved, joy may be elicited since the agent may now achieve its goal of "pursue other interests". The most important issue to comment upon here is that the "loss of job" event is neither desirable or undesirable, instead the consequences of that event are appraised in terms of the agent's goals and the desirability of the event is determined in context of that goal. This follows the OCC model's prescribed method of dealing with simultaneous opposing emotions unlike Steunebrink et al. who appear to propose their own method (see section 4.2.1).

The actions of agents are also modelled in much the same way i.e. there are two modal operators assigned to appraise the praiseworthiness/blameworthiness of an action in context of agent's current standards (which are internalised). It is also asserted by Adam et al. that a consequence of an action cannot be appraised as being both praiseworthy and blameworthy if appraised in context of the same standard.

The logical formalism of emotion eliciting conditions is somewhat more complex than those proposed by Stuenebrink et al. with Adam et al. also taking time into account. Yet, whilst Adam et al.'s formalism models the qualitative aspects of emotions in much the same way as Steunebrink et al., there is a distinct lack of quantitative aspects considered. This results in some important consequences of ignoring quantitative aspects the most important of which being: if the eliciting conditions for an emotion are met, then the emotion is always experienced. In other words, there is no gradual increase of an emotion's potential before that emotion is activated. This is an important point since it is particularly problematic with respect to subjectivity of emotional experience. Essentially, if the eliciting conditions for an emotion are the same for two distinct agents and those eliciting conditions occur then both agents will experience the emotion in question with equal intensity (due to the lack of quantitative measures of emotional intensity). Furthermore, the lack of consideration regarding quantitative aspects of emotions means that emotions remain active so long as their conditions stay true which, depending on the application domain, may not be at all intuitive or desirable. Finally, the logical formalism proposed by Adam et al. only takes into consideration the triggering conditions of emotions: it does not attempt to specify what effect the emotion's activation has upon the beliefs, desires or intentions of the agent. As such, the usefulness of the model only extends to the qualitative modelling of eliciting conditions for emotions and in its consideration of simultaneous opposing emotion experience.

4.3 Implemented Emotion Models

The research interests and motivations of computer scientists with respect to implementing or formalising emotional models can be quite diverse. Conveniently, Burghouts et al. identify three research motives that computer science researchers are driven by in [25]. The first, termed as the *cognitive-engineering* motive, focuses on how emotions may be used to provide immediate responses for computer systems that are situated in competitive, resource-bounded environments. Consequently, emotions are evaluated in terms of their pragmatic efficiency when utilised in this way. Those motivated by the *experimental-theoretical* motive are concerned with investigating whether or not particular computational emotional models are valid (do these computational models act in a faithful way when compared against the relevant psychological model) and how powerful the computational model is when it is used to explain real-world observations. The third motive, the *believable-agent* motive, describes research that is concerned with the implementation of computer systems capable of displaying emotion so as to suspend belief in human beings that such computer systems are machines.

The aim of this section is to present details and note important features of computational models of emotion and associated research authored by computer scientists. By doing this I hope to then be able to highlight the novelty of my own research and outline which features I will take forward. With respect to the three motives proposed in [25] I will also attempt to identify and consider research that may be clearly labelled as being driven by either the cognitive-engineering, experimental-theoretical or believable-agent motive. The intention of this is threefold: firstly, I will be able to clearly distinguish common themes between research motives; secondly, I may determine what my own research motive is; thirdly, I may contextualise my own research amongst the research considered. Details of the results of the papers discussed will not be presented since in the majority of cases the results obtained are not important with respect to this thesis.

4.3.1 Reilly

The computational model of emotion proposed by Reilly in [155] is focused upon implementing an interface for artists that is both easy to understand and use in order to allow for the creation and definition of believable emotional agents. Consequently, Reilly's research motivation may be classified as being a member of the believable-agent research motive proposed by Burghouts et al. in [25]. It is important to note that Reilly's emotional model is rather pragmatic in its approach and is designed to resolve complex implementation issues. As a result, some of the design and implementation decisions selected by Reilly may be adopted for my own work so as to facilitate a simpler implementation of the emotion modelling framework to be developed. Reilly's emotional model is predominantly based upon the OCC model (see 4.1.1.1) with the elicitation mechanism, emotions implemented, intensity mechanisms etc. all derived from work

presented in [142]. Since [155] is notably capacious in its scope, I will only consider relevant aspects of the computational model of emotion here.

The lifetime of an emotion in Reilly's model spans four stages, these are outlined below:

1. Generation of an emotion structure is triggered by an event creating an emotion structure.
2. The emotion structure is stored and decays over time according to a function defined by Reilly.
3. The current set of emotion structures is used to generate additional behavioural features which abstractly determine the actions of an agent.
4. Behavioural features are used to affect specific aspects of the agent's behaviour.

To generate an emotion a number of inputs are taken into account. I am only interested in a subset of these inputs namely: sense data/sensory memory, goals, standards and attitudes (Reilly defines these latter three concepts as Ortony et al. do in [142]). Sense data and sense memory take into account the agent's perception of its environment specifically: where artefacts are located, what other agents are doing etc. Such input is especially important to model agents that are emotionally reactive since this information forms the basis upon which an appraisal can be made. These appraisals result in an emotion and make use of the agent's internal goals, standards and attitudes (these concepts have been explained in detail in section 4.1.1.1 therefore I will not discuss them again here). Reilly uses the "Hap" language (originally designed to write robust, reactive physical behaviours for agents) to construct emotion generators. Some important features of this language are discussed below:

- Daemons: The ability to have rules that fire in certain circumstances is very important for creating emotion generators. Every generator Reilly constructs is expressed as a daemon that waits for a particular emotional situation to occur, and then fires, creating an emotion structure.
- Flexible match language: Used to code the left-hand sides of emotion generation rules (qualitative conditions used to describe the situations in which the generator should fire).

Therefore, each emotion could be said to be under the control of one daemon that appraises perceived events and matches relevant aspects to a set of conditions for that emotion using a flexible match language. If the conditions for an emotion are met, the daemon will generate the appropriate emotional structure for use. What is important about this is that the emotion generation procedure may be simplified to use a set of "if-else-then" conditions for emotional appraisals. Consequently, it is debatable whether an agent may experience opposing emotions simultaneously towards the same situation

or not. It should also be noted that, whilst this part of the emotion elicitation procedure appears to be qualitative, the daemon will convert the appraised event into a numerical value meaning that emotion generation is ultimately quantitative.

This quantitative approach to emotion elicitation facilitates the implementation of notions such as *emotion potentials* and *activation thresholds*. In Reilly's model, the potential of an emotion has to reach a certain threshold value before an associated behaviour is expressed in keeping with the OCC model, Frijda's model, Scherer's model and Steunebrink et al.'s logical formalism of emotion (see sections 4.1.1.1, 4.1.1.2, 4.1.1.4 and 4.2.1 respectively). Reilly appears to consider emotions as being functional since the elicitation of an emotion causes the expression of an un/intentional behaviour. This is not to say that Reilly does not view emotions in other ways e.g. generated emotion structures are capable of feeding back into the emotion generation system, but it is important to note that he does take a functional view regardless of this.

Reilly's simplification of eliciting conditions for emotions needs to be emphasised: based upon his emotional model, it may be sufficient for me to model an emotion such as anger by simply declaring that "an opponent has performed action α towards me, therefore my intensity of some emotion will increase/decrease". Such a rule is much easier to construct and utilise than a rule which fires if the agent disproves of another's blameworthy action along with it being displeased about the undesirability of the event. In this case, the agent would have to determine: what the action was, who performed it, the degree of intentionality behind the action, the consequence of the action and whether the consequence was desirable or not. Essentially, the appraisal process is greatly simplified.

Emotional effects are modelled by Reilly as being simple mappings of emotion to behavioural families that are pre-coded and stored in the agent's action repository. For example: anger could be mapped to aggressive behavioural features but may also be mapped to submissive behavioural features. Ultimately, it is the user of the system who decides these emotion-behaviour mappings but by allowing such flexibility, Reilly permits the creation of distinct personalities for distinct agents that use his emotional model. In addition, behaviours resulting from the elicitation of an emotion may or may not be triggered; an emotion's associated behaviour may therefore be masked. In other words, even though an emotion's potential may have surpassed its activation threshold, the behaviour that should result is suppressed (a landmark of Frijda's emotional model, see section 4.1.1.2). Unfortunately, Reilly does not appear to provide any details with respect to how this masking mechanism works so it is difficult to expound on this.

Reilly's emotional model is also notable since it makes an attempt to consider the issue of emotional decay (this issue was first encountered in the discussion of Steunebrink et al.'s logical formalism of emotion, see section 4.2.1). Like other aspects of his model, Reilly's emotional decay function is relatively simplistic, especially when compared with Steunebrink et al.'s method and is not based upon any solid psychological research. The function used normally reduces an emotion's intensity by 1 every tick, although for some

emotions (fear, for example), the current intensity of the emotion may be halved every tick.

4.3.2 "ACRES"

The *Artificial Concern Realisation System* or *ACRES* is a computational system that implements Frijda's emotional model (discussed in section 4.1.1.2). ACRES is presented by Frijda's student, Jaap Swagerman in [192] but I will restrict my consideration of the system to the more succinct description provided by both Frijda and Swagerman in [68]. The purpose of ACRES is to ascertain the validity of Frijda's concern realisation system described in section 4.1.1.2. Frijda and Swagerman's reasoning is such that if the preceding analysis of concern realisation is valid, it should be possible to construct a concern realisation system. Therefore, the research motivations of Frijda and Swagerman can be classified under the experimental-theoretical research motivation proposed in [25].

With Frijda and Swagerman's reasoning in mind, emotions in ACRES act functionally to safeguard concerns that provide the program with some independence in its interaction with its environment. In other words, if ACRES has a concern that is under threat, an emotion will be activated to remove the threat to that particular concern by identifying and implementing some intentional behaviour that is usually directed at the system itself or its operator. So, if ACRES current circumstances threaten its concerns then these circumstances should be avoided resulting in the system acting intentionally to increase the likelihood of its current circumstances being changed. Conversely, if ACRES current circumstances satisfy any of its concerns then the system should act so that the likelihood of current circumstances changing is decreased.

Concerns are represented by a theme and a tariff that represents the desired situation and what counts as an above-threshold undesirable/desirable state of affairs. This is an important point since concerns contain a notion of thresholds for realisation in much the same way as prescribed by the emotional models of Ortony et al. and Scherer (see sections 4.1.1.1 and 4.1.1.4 respectively) and the logical formalisms of Steunebrink et al. and Adam et al. (see sections 4.2.1 and 4.2.2 respectively). To illustrate how these thresholds are represented, Frijda and Swagerman consider the "preserve reasonable waiting times" concern. This concern is composed of two sub-concepts defining what a "long waiting time" and a "short waiting time" are (10 seconds and 2 seconds, for example).

To give an example of how the concern thresholds work imagine a situation where the current system's response time is greater than or equal to the time specified in the "long waiting time" sub-concept. In this case the "preserve reasonable waiting times" concern is realised and an emotion is elicited that identifies an intentional behaviour to reduce current and subsequent waiting times so that it is less than or equal to the value contained in the "long waiting time" sub-concept and greater than or equal to the value contained in the "short waiting time" sub-concept. From here the intentional behaviour identified may then be implemented by ACRES to remove the threat to the

relevant concern so long as the behaviour's expected consequences are admissible. To expand upon this point, Frijda and Swagerman state that the expected consequences of intentional behaviours are considered by ACRES before an intentional behaviour is performed. Therefore, it may be true that even though an emotion is elicited, its associated behaviour is not. Frijda and Swagerman use regulation processes to deal with this issue, but details of these processes are not provided.

Essentially, ACRES, for the most part is a reactive system since it only responds to user input but there are times when autonomous actions will be performed (such as asking a user to define an emotion). Generally though, there are only two ways in which an intentional action is initiated: user instruction or concern relevance (detected by the concern realisation functions). Furthermore, plans generated due to user instructions are not always executed: this occurs when another concern's relevance is detected and its above-threshold value is higher. This functionality is worth highlighting since it demonstrates the relationship between emotional intensity and action precedence: essentially, if one concern is under more of a threat than another or, if that concern is deemed to be more important, then the emotional response associated with the threatened or more important concern will be higher and candidate actions are prioritised. Therefore, the emotion produced by a concern being threatened is associated with a precedence ordering of actions of the agent.

4.3.3 "Cathexis"

The *Cathexis* system, designed by Velásquez, is noteworthy due to its innovative computational implementation of an appraisal model of emotion (see section 4.1.1). There is no monolithic psychological model of emotion underpinning Cathexis; the system appears to be amalgam of various noteworthy features borrowed from a range of psychological models of emotion. Cathexis is used by Velásquez in a number of simulations whose details are outlined in three papers: [207], [205] and [206]. Of these papers I focus on [205] since it provides a concise overview of the Cathexis system and an informative description of a simulation test-bed used to test the internals of the system. The purpose of this testing is to evaluate how useful Cathexis is with respect to enabling the creation of emotional agents therefore, Velásquez's research motivation can be considered to be experimental-theoretical.

Cathexis is a distributed, object-orientated system; emotions are modelled as networks composed of special proto-specialist agents. This design paradigm appears to be a literal translation of Minsky's concept of human intelligence outlined in [128]. In this work, Minsky proposes that intelligence is composed of interactions between "mindless" agents giving the work its title: "*The Society of Mind*". Each proto-specialist in Cathexis represents a basic emotion family; "basic" is to be interpreted as Ekman defines the term in [44] i.e. there are a number of separate emotions that differ from each other in important ways such as the events that elicit them, resulting bodily expressions, behavioural/physiological responses etc.

Every proto-specialist in Cathexis is endowed with multiple sensors that monitor both external and internal stimuli. External and internal stimuli that are relevant to a particular proto-specialist will either increase or decrease the intensity of the emotion family that the proto-specialist represents. Proto-specialists also run in parallel so multiple emotions can be experienced simultaneously by one agent. The concept of proto-specialists and their functionality is significant for two reasons: firstly, the idea draws parallels with notions that have been previously encountered such as Frijda and Swagerman's concerns (see section 4.3.2) and Reilly's demons (see section 4.3.1). Also, each emotion proto-specialist would appear to possess an associated intensity value which can change over the course of time (much like the emotional fluents defined by Steunebrink et al. in section 4.2.1). A further salient aspect of Cathexis is that each proto-specialist has three associated threshold values:

- α: controls the activation of the emotion. Once a proto-specialist's associated intensity surpasses this threshold, an output signal is released to other proto-specialists and Cathexis' behaviour system. This system then selects an appropriate behaviour according to the current state of the emotion systems.
- ω: specifies the maximum intensity for that emotion proto-specialist. This is consistent with real life emotional systems in which levels of arousal will not exceed certain limits.
- Ψ: the intensity decay function of the emotion.

These thresholds are again similar to concepts previously encountered i.e. the thresholds mentioned by Ortony et al., Scherer, Steunebrink et al. and Reilly (see sections 4.1.1.1, 4.1.1.4, 4.2.1 and 4.3.1 respectively) and the tariffs outlined by Frijda and Swagerman (see section 4.1.1.2). In addition, the ability of proto-specialists to be activated in parallel allows for various emotions/actions to be experienced/performed simultaneously and also allows for blends where two basic emotions may be elicited at the same time and the resulting emotional state of the agent is itself distinct (people who experience grief may be a blend of the emotions of anger and sadness, for example).

Each proto-specialist makes use of an intensity decay function that appears to copy Reilly's decay function (see section 4.3.1): the intensity of an emotion is decremented by 1 every time step until the emotion becomes inactive. This situation, where the decay function proposed is a "best-guess", is familiar since the decay functions proposed by Steunebrink et al. and Reilly (see sections 4.2.1 and 4.3.1 respectively) are also "best-guess" attempts[3].

As mentioned earlier, after a proto-specialist recognises that an emotion's intensity has surpassed its threshold of activation, an output signal is sent to the behaviour system. This system determines what behaviour is appropriate for the agent to display,

[3] "Best-guess" in the sense that none of the decay functions outlined are backed-up by psychological evidence obtained through experimentation

given its current emotional state. Again, the primary output of the emotional model detailed here is some intentional behaviour that is consistent with the emotional state of the agent. This would indicate that Velásquez aligns himself with a functional view of emotions i.e. an emotional state serves to alter some aspect of the current environment that is internal/external to the agent that the system is implemented in. Furthermore, behaviours are guaranteed to occur if the associated emotion is elicited, there is no concept of *action potentials* (see section 4.1.1.2) included in Cathexis.

"Simón", the simulation test-bed created and used by Velásquez in [205] consists of a synthetic agent that represents a young child which a user can interact with. Interactions provide stimuli to Simón which, along with the internal stimuli generated by his motivational systems, will cause him to react emotionally. The user can also manipulate the different parameters of Simón's proto-specialists; distinct configurations of these parameters enable the representation of different emotional reactions, moods and temperaments. Velásquez refers to these parameter configurations as the affective style of the agent. To put this another way, a user may create a number of distinct Simón agents, each with their own particular emotional *character*. Different versions of Simón with distinct emotional characters will make use of and produce different emotional inputs and outputs even though the basic components of the agent are the same. This is an important feature and will be taken forward in my own work.

4.3.4 "The Affective Reasoner"

Elliott's *Affective Reasoner* simulator is presented in a number of papers, specifically [46], [47] and [48]. Of these I will focus upon [47] as it is the most concise in its description of the system. In [47], the "Affective Reasoner" is populated with a number of agents each endowed with a set of appraisal frames that are akin to a unique emotional character much like Velásquez's concept of affective styles in "Simón" (see section 4.3.3). These appraisal frames represent an agent's individual goals, principles, preferences and moods i.e. it acts as a knowledge base for the agent. Different combinations of these appraisal frames can be used by the agents to interpret situations that unfold in the simulation consequently, these interpretations may or may not cause the eliciting conditions of particular emotions to be met. However, if an emotion is elicited, it is expressed using behavioural channels which can be perceived by other agents in the simulation and as new simulation events. Elliott therefore appears to view emotions functionally with some form of behaviour being guaranteed to occur when one is elicited. The Affective Reasoner therefore appears to have much in common with the Cathexis system (see section 4.3.3).

A key issue to address here is how Elliott implements emotion elicitation in the Affective Reasoner. The underlying theory used by the system is based predominantly on the basic idea proposed by the OCC model (see section 4.1.1.1). Domain-independent rules are used as a complement to the agent's appraisal frame to deal with emotion generation. So, when the contents of an appraisal frame match the left-hand side of

one of these rules, an emotion instance is generated. It is not explicitly stated as to whether or not opposing emotion instances may be generated given the same appraisal frame contents but it is reasonable to assume that this will not occur since appraisal frames appear to be quite specific with respect to the emotion generation rules. It would appear that intensity thresholds are not implemented in the Affective Reasoner so that emotion elicitation is not quantitative in nature but rather qualitative. Consequently, in context of the Affective Reasoner, an agent's interpretation of a situation is critical to the determination of how an emotion-eliciting situation maps into an eliciting condition.

In addition, the Affective Reasoner also makes use of so-called construal frames that represent the goals, principles, and preferences of an agent. These construal-frames echo Frijda's concept of concerns (see section 4.1.1.2). Therefore, when a situation occurs, an agent assesses elements of it according to its construal frame to determine if the situation requires further consideration.

To determine what behaviour should be expressed when a particular emotion is elicited, Elliott makes use of an action generation module. This module allows a further modelling of different emotional characters to occur as different behaviours may be mapped to different emotions. In practice, this computational modelling of character may be observed as the occurrence of one person shouting when they become angry and another crying when the same emotion is elicited.

Like Velásquez, Elliott's research motivation in [47] may be classified as experimental-theoretical since the purpose of the Affective Reasoner in this work is to explore the representational power of theory-based emotional architectures using agents. A simulation is used to test the validity of the emotional model. In these simulations a human user attempts to sell telephone book advertising in context of an interview/sales meeting to an agent.

4.3.5 "EBDI"

The *Emotional-Beliefs-Desire-Intentions* decision-making agent model is presented by Jiang et al. in [95] and uses primary and secondary emotions to augment the classic BDI decision-making model proposed by Bratman in [24]. "Primary" and "secondary" emotions in this context are intended to be defined as Damasio defines them in [35]: primary emotions are those emotions that are elicited when certain situations in the world occur i.e. I see a bear and am frightened (people do not have a default fear of bears but perceiving one should cause some kind of emotional reaction). Secondary emotions are learnt through experience i.e. if I am rude to another I may feel guilt or shame later. In context of the EBDI model, primary emotions are modelled and used to deal with the issue of bounded resources (see Simon [174]) by impacting upon the belief-revision, option generation and filter functions of the BDI model. Secondary emotions are used to refine decisions made using primary emotions (if time permits). The EBDI model is intended to include emotion in an agent model inspired by a human being's practical reasoning process. Essentially, Jiang et al. are interested in researching

what effects emotions may have upon the intentional behaviour of agents; consequently their research motivation may be described as being a part of the cognitive-engineering research motivation outlined in [25].

Jiang et al. model anger and gratitude; the importance of these emotions with respect to reciprocal behaviour was discussed in section 2.5 of this thesis. Anger and gratitude are represented and implemented in a simplistic manner: an emotion appears to be synonymous with a defined strategy of play that an agent can utilise given certain circumstances. Emotion-effect mapping is also one-to-one so, if an agent, y, lies to an emotional agent, x, then x will become angry with y and will subsequently decrease the priority of information received from y. If y tells the truth to x however, then x will experience gratitude towards y subsequently increasing the priority of information received from it. The effects of an emotion are therefore guaranteed and it would appear that opposing emotions may not be experienced with respect to the same event or agent. This is because an opponent may only either lie or tell the truth and each emotion appears to cancel the other when elicited.

It is important to note that, like Elliott's Affective Reasoner (see section 4.3.4), the EBDI model appears to model emotions qualitatively since it does not include any notion of discrete activation thresholds or intensity values. Unfortunately, this is not made explicit in [95] and so this conclusion is only an inference. Another important feature of EBDI is that emotions do not have an explicit effect on the intentional behaviour of agents. Rather, emotions exert an effect upon an agent's belief revision function when receiving information from other agents. Ergo, whilst Jiang et al. appear to consider emotions as being functional in an internal agent context, their effects are not directly functional in context of an agent's external environment.

Like Elliott's simulation described in section 4.3.4, Jiang et al. use their emotional decision-making model in the context of a MAS. However, in this case, all agents are autonomous; no human interaction is required. Agents play against each other in the context of *Tileworld*, a game designed and outlined by Pollack and Ringuette in [150] to help investigate the meta-level reasoning of agents.

The simulation aims to analyse how emotions can help to adapt individual behaviour in a MAS environment and whether or not emotions provide an increase in utility with respect to an individual agent. In [95], the authors compare the performance of emotional agents and so-called "rational" agents (of which there are two types: truthful agents who always tell the truth and selfish agents who always lie). The reason why these two terms are used to distinguish between the types of agents implemented is not made clear and as such it could be observed that Jiang at al. are proposing a dichotomy between the terms of "emotion" and "rationality". This supposed dichotomy is discussed in detail in chapter 3 of this thesis so I will not devote further attention to it here, although needless to say, I consider it to be false. The research motivations of Jiang et al. in this paper can be classified as being part of the cognitive-engineering motivation as the simulation

focuses on how emotions may be used to provide immediate responses for agents in a competitive, resource-bounded environment.

4.3.6 Oliveira

The MAS simulation implemented by Oliveira in [139] aims to analyse how emotions can be used to facilitate the adaptation of an agent's behaviour to that employed by others and whether or not emotion is capable of enhancing the monetary utility of individual agents. Consequently, the research motivations for this work can be classed as being a part of the cognitive-engineering research motivation proposed in [25]. A public goods game where pieces of a pie may be bid for is simulated (each agent may only bid for a maximum of three pieces) and, if the total bids from all agents are less than or equal to the total number of slices available, each agent receives the number of pieces they bid for. If the total number of pieces bid for by all agents is greater than the total number of pieces available then agents do not receive anything. As can be inferred from the game description, each agent's utility is equivalent to the number of pie pieces they receive at the end of the game.

The closing function and adaptation functions mentioned above are used in context of an automaton, defined by Oliveira as:

> "a decision rule consisting of a finite set of states, a transition function (which defines the transition between states) and a behavioural function (defining how the agent behaves in each state)".

An automaton therefore maps out the actions an agent can perform in its current state, the actions that may be performed in every possible state given the current state and the set of all states possible if a particular action is performed in a particular state. Oliveira notes that such automatons are difficult to compute if an agent posses limited or incomplete information about its environment (this echoes Simon's concept of "bounded rationality" [174]).

Consequently, an agent may reach an impasse in its decision-making if the automaton it has computed does not yield either a beneficial state or an action to achieve a beneficial state. In such a case, closing and adaptation functions are used: a closing function is used to close an automaton i.e. move the agent from one state to another whilst an adaptation function is used if an agent's best response (defined as behaviour that maximises the agent's individual total of discounted rewards) fails to achieve desirable equilibria. The particular behaviour exhibited when these functions are employed is dependent upon the emotions that the agent is endowed with.

Each agent in Oliveira's simulation is endowed with a maximum of two, distinct, emotions (which can be anger, apathy or patience) for its closing/adaptation function e.g. an agent may use anger for its closing function and apathy for its adaptation function. Therefore, rather than being modelled as distinct notions in their own right, emotions are modelled as strategies, much like Jiang et al.'s implemented emotional system discussed in section 4.3.5.

Oliveira's method of modelling emotions is particularly interesting since it does not appear to be based upon any general psychological emotional model. Instead, Oliveira models four emotions based upon research into their mechanisms conducted by independent parties, these emotions are:

- Anger: the agent attempts to minimize his opponents' rewards.
- Apathy: the agent always plays the same action independently of the rewards received.
- Patience: the agent attempts to reward his opponents if they change their behaviour in order to benefit the community.
- Confidence: the agent plays the same and does not adapt according to best response aiming to convince other player's to use his strategy; this gains the agent credibility.

In Oliveira's simulation, emotions are functional in that they directly enable agents to make decisions and perform intentional behaviour in context of competitive, resource-bounded environments. The intentional behaviour produced by these emotions is guaranteed to occur and may be focused upon both an agent and its opponents. This means that the concept of *action potentials* (see section 4.1.1.2) is again not considered in this emotional model. With regards to emotion-behaviour mapping it would appear that one emotion can be mapped to many behaviours as Oliveira himself stipulates when discussing the effects of the emotions modelled.

There are a few important points to note with respect to Oliveira's use of emotion in [139]. Firstly, "rational" behaviour is defined by Oliveira as the ability of an agent to calculate its best response in any given automaton. So, if an agent is not capable of calculating its best response then it may use an "emotional" strategy to close the automaton; emotion is used when rationality is ineffective. The implicit issue with respect to this is that Oliveira is proposing a dichotomy between rational and emotional behaviour in much the same way as Jiang et al. do (see section 4.3.5). Furthermore, only endowing an agent with one emotion for use with each function seems to produce quite inflexible behaviour. The agent's ability to compute closing and adaptation functions could be much more extensible if more than one emotion could be used with respect to the closing or adaptation function. Consequently, it can be asserted that opposing emotions may never be activated simultaneously in an agent since an agent only ever has one emotion available to it and there are no antonyms of the emotions modelled.

Furthermore, like the implemented emotional models discussed in sections 4.3.4 and 4.3.5, Oliveira's emotional model does not consider the notion of *potential* or *activation thresholds* for emotions. Instead, if a best-response cannot be calculated by an agent, the emotion that an agent has been endowed with for its closing function is used. The situation is identical when an agent uses an emotion in context of its adaptation function.

In this way, the intensity of the emotions implemented by Oliveira are constant and do not reflect a human-centric approach toward emotion modelling.

4.3.7 Bazzan and Bordini

Bazzan and Bordini's research into the integration of emotions in computer science is presented in [12] where they discuss the emotional model they have implemented, the details of the simulation constructed to investigate their research questions and the results of these simulations. Bazzan and Bordini use the OCC model (discussed in section 4.1.1.1) exclusively to model emotion. Their justification for this is that the model groups emotions according to their eliciting conditions which facilitates computational implementation since there are some pre-existing distinctions. Bazzan and Bordini model four emotions in [12], the names of these emotions, their eliciting conditions and effects upon intentional behaviour of agents are listed below:

- Joy: elicited if an agent has collected a certain number of points or if at least a certain number of neighbours are joyful. Elicitation of joy in an agent causes the agent to cooperate.
- Distress: elicited if an agent has not collected a certain number of points or if at least a certain number of neighbours are distressed. Elicitation of distress in an agent causes the agent to defect.
- Pity: elicited if a neighbour has not collected a certain number of points. Elicitation of pity in an agent causes the agent to cooperate.
- Anger: elicited if an agent has not collected a certain number of points and the opponent has defected. Elicitation of anger in an agent causes the agent to defect.

When an emotion is elicited in an agent, the target of this emotion is the agent's intentional behaviour which is altered so that an agent may increase its monetary utility; emotions are therefore functional. In [12], emotions are implemented as distinct values and their effects are manifest when this value (potential) has reached or surpassed a user-set threshold. These emotion potentials are intended to be calculated in much the same way as prescribed by the OCC model i.e. local and global intensity variables are taken into account and such variables are intended to be domain-dependent. Emotion modelling is therefore intended to be quantitative.

In the simulations used in [12], Bazzan and Bordini only endow agents with one emotion in the same vein as Oliveira (see section 4.3.6). Consequently, opposing emotions may not be experienced in context of the same emotion-eliciting event. Other non-emotional agents in the simulation may cooperate or defect depending upon the strategy they have been programmed with (details of these strategies are not provided in [12]). Furthermore, the percentage of initial cooperators in the population can be set by the operator. The simulations themselves take place on a two-dimensional, thirty by

thirty square grid with each agent occupying one square each (ninety agents exist in the environment at any one time). The boundaries of the environment are fixed and each agent plays against its eight immediate neighbours (four to the north, south, east and west of the agent and four to the north-west, north-east, south-east and south-west).

In each round, every agent will play a Prisoner's Dilemma game against its eight opponents with the highest scoring agent from the neighbourhood taking over the squares occupied by agents that scored lower. This style of play is intended to build upon work performed by Nowak and May [135] who also simulate an iterated Prisoner's Dilemma game where the strategies of successful players are propagated into neighbouring cells. The findings of Nowak and May's research illustrates that different equilibria will arise in different neighbourhoods as any one player does not play against the entire population. Therefore, Bazzan and Bordini wish to investigate if emotion may be able to disrupt this equilibria in neighbourhoods so as to promote cooperation in neighbourhoods of defectors. The research may then be described as being driven by the cognitive-engineering motive according to Burghout et al.'s classifications [25].

As Bazzan and Bordini note themselves in [12], their framework is intended to be generic so that it may be used as a starting point by those wanting to generalise the rules associated with the emotions implemented. Therefore, the functions responsible for calculating potentials of emotions are explained quite generally i.e. whilst the variables to be used in such calculations are listed, the calculations themselves are not. Another feature of the framework is that agents are guaranteed to perform the behaviour prescribed by an emotion if that emotion is elicited. This design choice echoes all other computational models of emotion discussed in this section with the exception of Reilly's emotional model (see section 4.3.1).

4.3.8 "SHAME+"

The *SHAME* and *SHAME+* systems that are presented respectively in [149] and [25] by Poel et al. and Burghouts et al. are examples of a MAS whose aim is to test the effectiveness of the emotional model developed. I will focus upon [25] since this paper provides details of the augmented version of the SHAME system. That is not to say that I will ignore [149] entirely, instead, if any references are made to this paper they will be indicated clearly to avoid any potential confusion between the two works.

Based upon the research goal outlined above, it can be asserted that the research motivation of Burghouts et al. is a part of the believable-agent motive since they wish to conduct research into how best to model emotion so that natural human interaction is possible (the authors explicitly place their work into this research motivation category themselves). Furthermore, the resulting social interactions between agents are investigated to consider the effects emotion may have upon this facet of an agent's operation. To achieve this goal, a model that relates emotional and cognitive processes with behaviour is required so that an agent can be perceived as performing naturally by way of expressing emotions and particular personality traits. To study the effect emotions

have on social interaction, some form of test-bed is required which incorporates such a feature; in [149] and [25] this is achieved by constructing a bespoke simulation.

Fourteen emotions are included in the SHAME and SHAME+ systems (since there are so many they will not be listed here) and are modelled as opposing pairs. Each pair shares one scalar value that represents the potential/intensity of the opposing emotions. By doing this, Burghouts et al. ensure that joy and distress, for example: cannot be active at the same time with respect to the same event as they are antipodes of the same scale. A positive (negative) intensity value on an emotion's scale corresponds to the positive (negative) emotion represented.

Emotion elicitation is triggered by an agent appraising an event in the simulation based upon methods described in the OCC model of emotion[4] (see section 4.1.1.1) but are not exact replicas. As inferred by the discussion of how emotions are implemented by Burghouts et al., emotion elicitation is quantitative and is driven by the occurrence of events that are external to the agent rather than events that are internal (homoeostatic events, for example). Simulation events are mapped to emotional states for each type of agent: therefore events are appraised by agents but this appraisal stops after identification of the type of event. Ergo, appraisal is not implemented exactly in the way envisioned by the OCC model [142] where an emotion is elicited by the agent assessing how an event may impact upon its goals, standards and attitudes. It would appear then that events cause a pre-determined positive/negative alteration of the emotion scalar it is intended to affect. So, an event such as "attacked by predator" may subtract ten from an agent's joy-distress scalar, for example.

For all emotions implemented in SHAME and SHAME+, *activation thresholds* are set to zero by default. Therefore, when an event alters the scalar value for an emotion pair, the positive/negative emotion represented by that scalar value is activated. The intensity of the emotion experienced is determined by calculating how far from zero the current emotion's potential has shifted e.g. if an agent's joy/distress scalar is altered positively by ten then the agent will experience joy with intensity of ten. There is an issue with respect to this method of implementation as it is extremely unlikely that an agent may be neutral with respect to an emotion pair (their shared emotional scalar must equal *exactly* zero for this to be true). An important feature of the emotional model implemented in SHAME and SHAME+ is that all emotional scalar values are bound at -100 and +100 so the intensity of any emotion may not continue to increase indefinitely. Essentially, such a feature implements the concept of a *saturation point* (first considered and implemented in Cathexis, see section 4.3.3) and is important to note as it reflects a realistic, human-centric emotional model.

The SHAME and SHAME+ systems are most notable with respect to this thesis by virtue of their consideration of emotional *personalities*. In context of SHAME+, personality defines emotional characteristics i.e. what emotions an agent can experience, how strongly/weakly they are experienced when elicited and how long the emotion lasts

[4]Details of the appraisal process are taken from [149] where they are described in more detail than in [25].

for when elicited. The definitions for "strong" and "weak" are ambiguous in this case; it may mean that the threshold for an emotion's activation is lower or it may mean that a certain event type causes a greater shift in emotion intensity. To the best of my knowledge, these definitions are not clarified in either [149] or [25].

Turning to the issue of emotional effect it should first be highlighted that an emotion affects an agent's intentional behaviour and the intensity of an agent's emotions does not affect the behaviour exhibited. The question may then be asked, what is the purpose in implementing emotional intensities if they result in non-variance of intentional behaviour? This question may be answered by considering the decay function utilised by the SHAME and SHAME+ systems which makes use of knowledge provided by the user to alter an emotion's intensity over a period of time. Unfortunately, unlike Stuenebrink et al.'s, Reilly's and Velásquez's emotional decay functions (see sections 4.2.1, 4.3.1 and 4.3.3 respectively), details of how the function operates are not discussed. It is therefore difficult to assess whether this function could be adopted/modified for my own purposes or if it has any sound psychological foundation.

Burghouts et al. also model "self-control" which removes or reduces the dissonance between expected behaviour and behavioural standards. The mechanism used for this step is inspired by Bandura's observation [8] that emotional self-control and the potential actions that may result from particular emotions being elicited includes self-monitoring via. personal standards and corrective self reactions. Therefore, the emotion's effect cannot be said to be entirely guaranteed as this self-control may suppress the associated action. Justification for including a mechanism for self-control is provided by Gifford [96] who states that self-control is required to reduce conflicts between present and future gratification. The calculations and functions for applying self-control are not outlined in [25] therefore, it is difficult to ascertain how this should work and consequently, I am not able to adopt/modify such a function for this thesis.

4.3.9 "EPBDI"

Zoumpoulaki et al. introduce the *Emotion, Personality, Beliefs, Desires and Intentions* model (*EPBDI*) in [216]. Note that this model is not a direct extension to the EBDI model proposed by Jiang et al. in [95] (see section 4.3.5). The important difference between EBDI and EPBDI is the concept of *personality* which was first mentioned in section 4.3.8. Zoumpoulaki et al. computationally model personalities using the *OCEAN* model (discussed below) whilst emotions are modelled using the principles set out by the OCC model [142]. In [216] it is proposed that, by modelling personality and emotion, the required mechanisms for simulating realistic human-like behaviour under evacuation can be outlined and verified.

To this end, Zoumpoulaki et al. implement the following emotion pairs using the OCC model as a basis: joy/distress, hope/fear, pride/shame, admiration/reproach and sorryFor/happyFor (first three pairs affect the agent that experiences the emotions, the last two affect other agents). Given the discussion in [216], these emotions appear to be

modelled in much the same way as emotions are modelled in the SHAME and SHAME+ systems (see section 4.3.8) but is neither confirmed or denied by Zoumpoulaki et al. Since emotion pairs are modelled using scalar values (as prescribed by the OCC model) it would appear that emotions are modelled quantitatively in [216].

The emotions modelled act functionally to affect the perception of an agent, the agent's appraisal of events, an agent's decision-making and the actions performed by an agent in a given situation to cope with that situation in a particular way. The consequences of specific emotions upon these components along with whether or not the effects of emotions are guaranteed is unfortunately not made clear. Despite this, it is important to remember that emotion has some effect upon the agent's decision-making and the intentional actions subsequently performed. Since emotions are modelled as opposing pairs and appear to be driven by events/appraisals, it would be reasonable to assume that opposing emotions may not be active with respect to the same eliciting condition although confirmation of this is not made explicit in [216].

An agent's personality affects how the agent appraises events and its decision-making that results from this. As stated earlier, an agent's personality is modelled using OCEAN, an acronym of characteristics that stands for: openness, conscientiousness, extraversion, agreeableness and neuroticism. However, like the implementation of emotions discussed previously, insufficient implementation details are given in [216] with regards to how personality influences appraisal. Therefore, these ideas have not been implemented in the model used in this thesis.

Zoumpoulaki et al. make use of a simulation to investigate whether computational modelling of emotion produces more "human" or believable behaviour. Therefore, Zoumpoulaki et al.'s research motivation may be classed as being a member of the believable-agent research motive [25]. In this simulation, agents are situated in a two-dimensional environment consisting of static objects and randomly generated pockets of fire that may hurt or kill agents. The initial agent population's demographic/personality distribution and the position/spread parameters of fires are all user-defined.

4.4 Chapter Summary

In this chapter I outlined the various ways in which emotions could be modelled by first considering a number of psychological emotion models in section 4.1. Following this I then discussed prominent logical formalisms of some psychological models of emotion (see section 4.2) before presenting a review of literature concerned with how psychological models of emotion have been modelled computationally and used to answer various research questions (see section 4.3).

The literature reviewed in this chapter has led to a number of important conclusions that will serve to inform and direct my own emotional modelling framework. In chapter 5 I will use the conclusions drawn from the discussions presented in sections 4.1, 4.2 and 4.3 above to construct my own research agenda and how I intend to answer the research

questions posed. Specifically, this will entail a description of the emotional model I have developed and a discussion of the test-bed that I have used to enable an investigation into the research agenda proposed. Table 4.1 summarises the key points of the research discussed in section 4.3 that are of interest to this thesis. The intention of this table is to facilitate a comparison between key features of this research and my own emotion modelling framework that is outlined in chapter 5.

With respect to these conclusions a number of salient observations have been made. The first is that appraisal models of emotion (see section 4.1.1) are particularly well suited as a basis upon which computational models of emotion can be developed. Of the three types of emotional models considered (appraisal, dimensional and anatomical), appraisal models are the only type of emotional model to both consider the origins of emotion and consider them from a non-physiological perspective. Establishing how emotions are elicited is particularly important since without such knowledge, emotions themselves may never be modelled. The non-physiological standpoint is also exceptionally beneficial as computers are obviously incapable of physiological responses. Particular attention should be afforded to the OCC model of emotion (see section 4.1.1.1) since this appraisal model was developed with computational implementations in mind. This makes the OCC model especially tractable and attractive from a computational standpoint.

Aside from these observations, the review of literature in this chapter has highlighted an important niché in which my own work can exploit so as to produce novel, interesting research. The majority of research presented in section 4.3 is motivated by cognitive-engineering research proposed by Burghouts et al. in [25]. This field of research is quite interesting since there are many avenues of enquiry still unexplored. As Bazzan and Bordini state in [12]:

> "..little work has focused on the investigation of interactions among social agents whose actions are somehow influenced by their current emotional setting."

My own research will therefore be conducted in the context of this motivation as there are good foundations for the design and implementation of such research whilst there are still ample opportunities to produce novel work.

A succinct summary of the salient observations considered in this chapter can be found below.

- The most researched and appropriate general psychological emotional model for use in computer science appears to be the *appraisal* model of emotion.
- Logical formalisms have focused upon *appraisal models*, most notably the OCC model. Five of the nine implemented computational emotion models discussed are also based upon the *OCC* model.

TABLE 4.1: Key points of research discussed in section 4.3.

Research	**Public Goods Testbed?**	**Psych. Model Used**	**Emo. Elicit Quan./Qual.?**	**Emo. Pot. & Act. Thresh.?**	**Multi. Emo. per Agent?**	**Emo. Modelled**	**Emo. Affect**	**Emo. Ch.'s?**	**Emo. Decay?**
[155] Sec 4.3.1	No	[142]	Quan.	Yes	Yes	OCC's 22 + resentment, frustration, startle	Unintentional & intentional behaviour	No	Yes
[68] 4.3.2	No	[67]	Quan. & Qual.	Yes	Yes	Unknown	Intentional behaviour	No	No
[205] 4.3.3	No	Many inc. [128], [44]	Quan.	Yes	Yes	Unknown	Unintentional & intentional behaviour	Yes	Yes
[47] 4.3.4	No	[142]	Qual.	No	Yes	OCC's 22 + liking, disliking	Many (see fig 3 of [47])	No	No
[95] 4.3.5	Yes	[24], [35]	Qual.	No	Yes	Anger & gratitude	Priority ordering of info from opponents	No	No
[139] 4.3.6	Yes	None	Qual.	No	Yes	Anger, apathy, patience, confidence	Intentional behaviour	No	No
[12] 4.3.7	Yes	[142]	Quan.	Yes	No	Joy, distress, pity, anger	Intentional behaviour	No	No
[25] 4.3.8	Yes	[142]	Quan.	Yes	Yes	14 (see [25])	Intentional behaviour	Yes	Yes
[216] 4.3.9	Yes	[142]	Unknown	Unknown	Yes	10 (see [216])	Unintentional and intentional behaviour	No	No

- *Simplifying* computational models of emotion is sometimes preferable if the emotional model developed is to not be used in a variety of contexts.
- Emotions are typically elicited though a consideration of how the *consequences* of events or actions may affect an agent's *goals*.
- Typically, eliciting conditions are modelled *qualitatively* whilst their effects upon emotional potentials are typically modelled *quantitatively*.
- In the majority of cases, emotions are modelled to be *functional* in that they alter the *decision-making* or *intentional behaviour* so that the current situation may be maintained or altered.
- In computer science, investigation of research questions using emotional models is often conducted in the context of *competitive MAS simulations*.
- In the majority of cases, when simulations are used, *initial conditions* can be set by the user to explore either how initial conditions *affect* the emotions produced in agents or how emotions *enable* agents to deal with the initial conditions set.
- The research considered in this chapter introduces the notion of developing emotional models that enable the implementation of different emotional *characters* and *multiple emotions per agent*. This will be significantly expanded in my own research.
- Research classed as *cognitive-engineering* is relatively *unexplored* at the time of writing and so has ample scope for developing novel research.

Chapter 5

Research Agenda, Test-Bed and the Emotion Model

After considering the existing research presented in chapter 4 the context in which the rest of this thesis will be embodied can now be proposed. This context and consequently, this chapter, broadly consists of four sections: section 5.1 delineates the research questions I wish to answer, section 5.2 describes the test-bed that will be used to answer these questions, section 5.3 lays bare the details of the emotional model that forms the framework used to functionally model emotions so that they may be embodied in the agents that populate the test-bed, and section 5.4 defines my concept of emotional characters. In presenting these sections I hope to further contextualise and highlight the novel aspects of my own work when compared with other relevant pieces of research in previous chapters. Note that the information presented within this chapter is intended to be quite general since subsequent chapters will go into specific detail with respect to the sections outlined.

5.1 Research Questions

The main intention of this thesis is to investigate how emotion affects the interactions between social agents that play public goods games in the context of a MAS. The decision to tackle this research question was originally inspired by Bazzan and Bordini's observation (quoted in section 4.4 of chapter 4) that:

> "...little work has focused on the investigation of interactions among social agents whose actions are somehow influenced by their current emotional setting."

Although some attention has been paid to the topic recently, [95, 112, 113, 139], there is still much to investigate. Furthermore, the research presented and discussed in chapter 2 of this thesis led me to focus upon the effects that particular emotions have upon interactions between humans in the context of public goods games. To square

this with Bazzan and Bordini's observation it can be seen how "interactions" denote the "interactions among social agents" identified by Bazzan and Bordini and "emotion" can be translated as Bazzan and Bordini's concept of "current emotional setting".

Consideration of the concept of emotional characters (first encountered in section 4.3 of chapter 4) augments the initial research question proposed so I may now pose the following question: "*how do particular emotional characters affect interactions between social agents that play public goods games in a MAS context and why?*". In [25], Burghouts et al. show the effects of emotional character upon intentional behaviour by using a simulation where agents interact within a predator-prey environment. To survive (the main goal of each agent in the simulation), predator and prey both require food, water and health; if any of these values fall to a certain level, the agent will die. Two emotional characters are implemented: "Hero" and "Grumph", and agents may form groups. When the "Hero" character is the leader of the group the following behaviours are noted with respect to the "Hero" and "Grumph" emotional characters:

- "Grumph" only attacks predators if "Hero" has already attacked a predator.
- The first resources are taken by "Hero" with "Grumph" receiving resources later in the simulation. Eventually, the resources seem to be equally divided.
- Normally, "Grumph" only joins a group once.
- If the health of "Hero" decreases significantly and "Hero" instructs the group to search for a resource "Grumph" does not want, "Grumph" leaves the group and immediately joins another.

Despite these observations, Burghouts et al. do not provide any discussion of previous research to justify why these behaviours are produced by these emotional characters in [25]. In my opinion, this is an important point to comment upon since stating what behaviours are entailed by particular emotional characters is only half an answer. Therefore, in this thesis I seek to present observations of behaviour with respect to the interactions between agents with different emotional characters and also to explore in some detail why these interactions are produced by the emotional characters implemented.

Additionally, emotional characters may be pitted against each other to determine which is more "successful" in a particular context. Obviously, the term "success" could be defined in a number of different ways since its definition is context-dependent therefore, I will abstain from clarifying its meaning here and will instead provide specific definitions in subsequent chapters. From here, I may investigate whether the success of an emotional character is independent of the current game environment or whether different game environments select for the success of particular emotional characters. Understanding the reasons behind these outcomes will allow agent developers to make informed decisions regarding the emotional characters that could be implemented in

agents given particular situations. It may also help to develop a sociological understanding of why people with particular emotional characters thrive or languish in different contexts. Most importantly however, determination of whether emotional character success is affected by particular game conditions allows me to provide a much more complete answer to the research question posed in section 1.1 of chapter 1.

Given the research questions posed above, some other fundamental concerns also need to be addressed:

1. How should the emotions considered be modelled/implemented and why?
2. How should emotional characters be modelled/implemented?

Since emotional experience is inherently subjective, the question of how an emotion should be modelled is open to attack from several directions. To reduce the number of possible disagreements that may arise it is important to ensure that the emotions to be modelled and implemented are done so in environments that are capable of limiting the number of eliciting conditions and effects of particular emotions. Furthermore, these eliciting conditions and effects should be grounded in rigorous psychological research, otherwise debates based upon subjectivity of emotional experience could occur indefinitely. I will refrain from commenting upon how and why particular emotions will be implemented in this section since questions such as this will be tackled in subsequent chapters. The question of how emotional characters should be implemented will be also addressed in later chapters of this thesis but providing a solid foundation upon which others may build their work would be extremely beneficial for other researchers in the field of affective computing.

With respect to answering how emotions affect social interactions in public goods games, there are very few pieces of research that touch upon the subject (as stated by Bazzan and Bordini in the introduction to this section). Of the work discussed in section 4.3 of chapter 4 the following pieces of research are concerned with such research. For each relevant paper I have briefly summarised the main findings so that I may identify other possible research questions or compare relevant results with my own in later chapters:

- Jiang, Vidal and Huhns in [95] observe that, when compared with non-emotional agents, emotional agents always perform better since their behaviour is adaptive to the dynamic environment of Tileworld. However, no explanation as to *why* this result is observed other than that the emotional agent is more adaptive to the dynamic environment used.
- Oliveira in [139] demonstrates that, in the context of the pie-sharing public goods game, emotions do indeed help to increase an agent's individual score. Oliveira notes the following with respect to the emotions implemented:
  - Patience - performs well as an adaptation function.

  - Anger - performs poorly as an adaptation function unless opponents are patient and accommodate the agent's demands. However, anger works well as a closing function since the threat of punishment coerces opponents to stick to the equilibrium.
  - Apathy - performs best as a closing function since it almost always dominates anger and patience (except in the case of ten pieces of pie when the best adaptation is achieved by anger).
  - Confidence - an optimum level of confidence is observed: too little and agents adapt behaviour too quickly, too much and agents never adapt their behaviour; both strategies lead to a reduction in individual utility.

- Burghouts et. al. [25] make a number of observations regarding the behaviour of the "Hero" and "Grumph" agents implemented in their simulations so many in fact that they will not be discussed here in the interests of brevity and redundancy. The problem however, is that there is very little discussion of *why* these results are observed in context of particular emotions i.e. it is noted that when "Hero" is the leader of a group, "Grumph" is less likely to attack a predator if "Hero" has already attacked but what emotions cause this observation is no mentioned. Consequently, it is difficult to generalise these results outside of the context of the simulation environment used, all results pertain to an extremely niche simulation set-up which bears little similarity to the more common, well-known and generalisable public goods games discussed in section 2.1 of chapter 2.

- Bazzan and Bordini [12] find that the percentage of cooperators in the population can reach 47% if some agents are endowed with emotions and further experiments varied the percentage of initial cooperators but no significant change in the results were found. It was also observed in all simulations run that a clustering of cooperators and defectors existed i.e. like-minded behaviour tended to be proximally close. However, there is no explanation in [12] with respect to *why* the emotions implemented produce these results which is quite disappointing since such an explanation would be both useful and interesting. This is therefore an avenue whose novelty can be exploited in this thesis.

To summarise, I am interested in providing answers to the following questions in the remainder of this thesis so that the general research question posed in 1.1 of chapter 1 can be answered fully:

1. *What* emotions should be modelled computationally in the public goods game considered?

2. *How* should these emotions be modelled?

   - What *eliciting conditions* should be taken into consideration?

- What *effects* should the emotion have upon the agent's intentional behaviour in context of the current environment?

3. *Why* should these emotions be modelled in the manner described?
4. *How* should emotional characters be computationally modelled and implemented?
5. *Do* different emotional characters affect interactions between social agents in context of a public goods game and how?
6. *Why* do these emotional characters affect interactions between social agents in context of a public goods game in this way?
7. Do different game contexts have any affect upon the success of implemented emotional characters i.e. is the success of emotional characters *context-independent* or *context-dependent*?

The research questions proposed would classify my research motivation as being part of the cognitive-engineering research motive described by Burghouts et al. in [25]. This is primarily due to the general aim of this thesis which is to investigate how emotions may be used to provide responses for agents that are situated in competitive environments. Research questions 1, 2, 3 and 4 may also cause the work in this thesis to be considered as part of the experimental-theoretical research motive [25]. These questions are concerned with the validity of the approaches proposed by myself in relation to the psychological theories of emotion used in this thesis. Additionally, the emotional system that is to be implemented will be used as an experimental environment to verify the validity of the methods used. This adds further support to the classification of some of my research interests as being experimental-theoretical. As mentioned in chapter 4, my research interests in this thesis may not be classified as being a part of the believable-agent research motive since I am not concerned with encouraging the belief that the agents implemented are more than machines through the expression of emotion by these agents.

5.2 Simulation Test-Bed

To gather data that will be used to answer the research questions posed in section 5.1 I will implement a simulation test-bed whose details will be described in this section. The decision to use a simulation was inspired by the use of simulations in every piece of research discussed in section 4.3 of chapter 4 and also by a number of researchers who were mentioned in chapter 2, especially Axelrod ([7]) and Nowak et al. ([136], [137] and [134]). The advantages of simulation are numerous, especially in the context of this thesis since simulation permits the control of all necessary game parameters and also allows the harvesting of large amounts of data in a relatively short time. It is also easier to explain observed behaviour in the context of a "real" environment than it is in a purely abstract, analytic setting. Finally, since emotions and their various facets are

inherently probabilistic, using simulations enables a wide-range of these probabilities to emerge (so long as enough repeats of a simulation are run) providing a more valid set of results and therefore conclusions to be produced.

As can be seen in section 4.3 of chapter 4, most simulations that are concerned with emotional models make use of competitive MAS games. Inspired by this, I have developed a competitive MAS simulation based upon the iterated Prisoner's Dilemma game. The iterated version of the Prisoner's Dilemma is a competitive, public goods game where the rational action for agents is to defect despite the chance for opponents to retaliate in subsequent rounds following a defection (see the "backward induction paradox", discussed in section 2.1.2 of chapter 2). Since I would like to investigate how emotion can enable cooperation between agents as well as investigating how emotion affects societal interactions in general, the iterated Prisoner's Dilemma provides a good context.

Further to this, many psychological studies designed to investigate the emergence of cooperation between human players make use of the iterated Prisoner's Dilemma to collect data enabling their research questions to be answered (see sections 2.3, 2.4, 2.5 of chapter 2). The fact that the game has been used so frequently and so broadly by psychologists facilitates answering questions such as "*what emotions should be modelled?*" and "*what are the emotion's eliciting conditions/effects?*".

The widespread use of the game is most likely due to both the pessimistic rational prediction of the game's outcome (see above) and because there are very few variables in the game that can influence outcomes. This ensures that any observations or results gleaned from using the game as a test-bed are relatively free from contamination by unmanaged variables. For example: players have a limited number of possible intentional actions (cooperate/defect) which makes emotion elicitation and the effects of emotion easier to model. As a consequence of these limited actions there are only four possible outcomes for each round played: CC, CD, DC, DD. Each outcome is clearly distinguished from the others so modelling emotion elicitation is again facilitated. The pay-off matrix for the Prisoner's Dilemma is implemented in the simulation as shown in table 2.1 of section 2.1.2.

The iterated variant of the Prisoner's Dilemma game permits me to analyse how social interactions between agents develop over a period of time. Using the standard one-shot version of the game would not yield such information rendering some of the research questions posed unanswerable. Further to this, the one-shot version of the game is not suitable since my agents require the ability to evaluate events and then modify their emotional state and intentional behaviour in light of them. Therefore, since agents need to be able to react to the actions of others and build relationships, the iterated version of the game is preferable. Using the Prisoner's Dilemma also enables me to reuse algorithms devised by previous researchers for my simulations, reducing problems of experimental bias.

FIGURE 5.1: Graphical representation of player/comparator subset divisions.

5.2.1 Implementation Details

Simulations are implemented using "NetLogo" [210], further details are available in [112] and [113]. The simulation environment is a two-dimensional space and was inspired by the same common design choice employed by researchers such as Jiang et al., Bazzan and Bordini, Poel/Burghouts et al. and Zoumpoulaki et al. (see sections 4.3.5, 4.3.7, 4.3.8 and 4.3.9 of chapter 4 respectively). In these pieces of work, the simulation environment is also composed of a two-dimensional space upon which a number of agents are situated. The size of the space implemented in my own simulations varies depending upon the version of the simulation being run (details of this are provided in subsequent chapters). It should be noted that the simulation environment does not have *edges* except for the first version of the simulation, discussed in section 6.4 of chapter 6. If Bazzan and Bordini's simulations in [12] are considered then the disadvantages of imposing edges become clear. Firstly, agents situated close to edges have fewer opponents than those located further away. Also, if the success of an agent is determined by its proliferation through a population then edges could dramatically affect the results obtained. For example: whilst some successful agents may propagate themselves into a maximum of eight neighbouring spaces at any one time, agents located close to edges may only be able to propagate into a maximum of five. Thus, it may be that some successful strategies are deemed unsuccessful for no other reason than they are played by agents whose propagation is spatially limited compared to other agents in the environment. This point is especially important when considering the emotions and simulation environments modelled in chapters 7 and 8.

Depending upon the simulation version used, agents may have up to two sets of opponents depending upon the simulation context. These sets are named respectively as the *player* and *comparator* sets. The player set represents the set of agents that an agent plays the Prisoner's Dilemma against in each round. The comparator set represents the set of agents that an agent compares its score against when the *comparison round* is reached. The value of the comparison round is specific to particular versions of the simulation and is therefore discussed in subsequent chapters. Player and comparator sets may be further divided into four subsets: *parallel*, *orthogonal*, *diagonal* and *octagonal* sets. The agents that constitute these sets are illustrated in figure 5.1 for clarification purposes.

The blue agent in figure 5.1 denotes the agent whose player/comparator set is being described in context of. Yellow agents represent those that may comprise a parallel player/comparator set, whilst a combination of the red and yellow agents denote an orthogonal player/comparator set. Green agents illustrate the agents that comprise a diagonal player/comparator set and a combination of the yellow, red and green agents represent an octagonal player/comparator set.

The subset of agents that constitute an agent's player and comparator sets may be different, therefore it may be the case that whilst an agent's player set is composed of the parallel agent set, its comparator set may be composed of the orthogonal agent set. Player/comparator subsets are implemented simulation-wide i.e. it can never be the case that one agent has a player set made up of parallel agents and another has a player set composed of orthogonal agents. The same is true for comparator sets; one agent may not have a comparator set made up of parallel agents whilst another has a comparator set made up of orthogonal agents. Varying the configurations of these sets enables variations in the number of, and overlap between, the opponents played; these results are then used to compare performance of agents in different contexts.

The simulation environment houses a number of agents with different strategies or emotional characters. Agents are capable of two basic actions: cooperation or defection and mobility is restricted, no matter the particular simulation version being used. At the conclusion of each round, each agent determines its pay-off according to its own behaviour in the last round and each of their opponents. Pay-offs are calculated as per the Prisoner's Dilemma pay-off matrix shown in table 2.1 of chapter 2.

For the simulations discussed in chapters 7 and 8 certain initial parameters for each agent present in the simulations are randomly allocated by the underlying simulation code on a per-agent basis. These initial parameters are:

- An agent's initial behaviour of i.e. does an agent cooperate or defect initially? Relevant for simulations used in chapters 7 and 8.
- An agent's initial emotional character. Relevant for simulations used in chapter 7.
- Location of each agent in the simulation space. Relevant for simulations used in chapters 7 and 8.

Finally, the number of rounds to be played in any of the simulations run is never made known to any of the agents playing. This does not really matter since agents cannot do anything with this information unless they are programmed to do so, but in the interests of completeness I will make this feature of the simulations clear here.

5.2.2 Simulation Progression

The aim of this section is to make clear the general order of events that occurs in every version of the simulation implemented in this thesis. As with the size of the simulation

environment, the specifics of some steps may differ slightly for particular simulation versions but the following basic order is always adhered to:

1. All agents including the mediator are placed in the environment. These placements are pre-determined and are constant throughout the course of the simulation although the placing of agents is version-specific in context of the simulation.

2. Agents then choose to cooperate or defect and their action is sent to their opponent(s).

3. Each agent's emotional state and individual score is then updated accordingly (specific details for each simulation version are discussed in subsequent chapters) and if an emotion's potential has surpassed its threshold then the agent may display the associated intentional behaviour in the next round.

4. The next round is played until the round limit is reached.

In section 2.3 the role of social interaction prior to a Prisoner's Dilemma game, which has been explored in some of the experiments carried out with human subjects [64, 65, 162], was discussed. In the simulations run, however, no agents engage in a "conversational phase" prior to a game beginning. Thus, no emotions are elicited before the game begins and emotions instead develop during run-time. In effect, the early rounds are a phase in which agents "get to know one another" and adopt emotional attitudes to one another, which then typically harden into stable patterns of interaction, as explored in chapter 8. Because these initial turns are where agents form their emotional relationships, it could be seen as analogous to a conversational phase.

5.3 The Emotion Model

After considering other implemented computational models of emotion in section 4.3 of chapter 4 it is clear that there is no one definitive computational model of emotion. Instead, various computational models of emotion have been proposed that appear to focus on distinct aspects of the emotion elicitation/effect process depending upon the aspect intended to be investigated. In accordance with this I have developed and implemented an emotional model that considers emotions to be *functional* in accordance with Ortony et al., Frijda and Scherer (see sections 4.1.1.1, 4.1.1.2 and 4.1.1.4 of chapter 4). That is, I focus on the functional role of emotions with respect to the decision-making concerned with selection of an agent's intentional behaviour. Whilst I acknowledge that emotions can have physiological and psychological effects I do not consider such aspects in this thesis. In this section I will discuss the emotional model that has been developed in detail, although I will postpone in-depth discussions of the particular eliciting conditions and effects of particular emotions until their occurrence in subsequent chapters. This emotional model forms the framework used to functionally model emotions so that they may be embodied in the agents that populate the test-bed described in section 5.2.

The emotional model used in this thesis is based predominantly upon the OCC model of emotion [142] for the advantages it imparts (discussed in section 4.1.1.1 of chapter 4) and is relatively simple. For example: an agent does not actually appraise events in the simulation; instead it is assumed that such appraisals are universal and a particular event triggers a pre-defined emotional response. All agents are therefore homogeneous in their emotional appraisals and reactions: what differs between agents is how slowly or quickly an emotion is elicited (this discussion strays into the area of emotional characters which is subsequently addressed in detail in section 5.4 of this chapter). The decision to develop a simplified computational model of emotion was inspired by Reilly (see chapter 4, section 4.3.1) since the computational model of emotion proposed in [155] is quite pragmatic but still suffices to allow for investigations into research questions relevant to his work. Creating a pragmatic emotional model confers a number of advantages; ease of implementation and therefore validation are perhaps the most notable benefits. Due to this simplification of emotional modelling, I am able focus my efforts upon result retrieval and analysis.

There are four emotions modelled and implemented in this thesis; their exact effects upon the intentional behaviour of agents is analysed in the next three chapters. A list of these emotions and the relevant chapters they are discussed in can be found below:

- Anger: discussed in chapter 6.
- Gratitude: discussed in chapter 6.
- Admiration: discussed in chapter 7.
- Hope: discussed in chapter 8.

At this point I will introduce my definitions for the terms "emotion instance" and "emotional class" since they have specific meanings in context of this thesis. As an object is a singular instance of a class in object-orientated programming terminology so is an emotional instance an instance of an emotional class. The inspiration for using such terminology is taken from Elliott who also defines emotional instances in this way in [46]. Therefore, in the emotional model proposed, agents may be endowed with multiple emotional instances of the same emotion class and may also be endowed with more than one emotion class.

The maximum number of emotion classes implemented in any one agent is *three*: anger and gratitude along with either admiration or hope. The number of emotion instances implemented in any agent is equal to the number of opponents an agent faces. Therefore, the maximum number of emotion instances present in any given agent is *eight* per emotion class since the maximum number of opponents any agent may face is eight (an octagonal player set). Implementing emotions in this fashion allows for richer societal interactions due to increased behavioural flexibility which in turn will provide richer data sets for analysis. In contrast, researchers such as Oliveira [139] and Bazzan and Bordini [12] (discussed in sections 4.3.6 and 4.3.7 of chapter 4 respectively) are limited in their

approach to this issue. The agents implemented in their simulations are only endowed with one emotion class each and it is unclear whether their agents have discrete emotion instances for each opponent or whether agents have one generic emotional state that informs their decision-making with respect to all opponents. Therefore, my goal is to build upon their observations by endowing the agents in my simulation with multiple emotions.

It is important to remember that specifics of the emotion model listed here are not intended to be directly reused by others (unless they too are using simulations that make use of the iterated Prisoner's Dilemma). Since I have focused upon modelling aspects that are relevant to the research questions posed in section 5.1, the specifics of the emotional model presented are *context-dependent*. So, whilst the general principles of the emotional model used in this thesis may be reused (emotions have eliciting conditions, effects etc.), careful consideration of its specifics should be afforded before their reuse. For example: in this thesis, the eliciting condition for anger is defection and eliciting anger causes an agent itself to defect. Before reusing such specifics one would need to question if such design choices are suitable for the context in which an emotional model is to be implemented. Therefore, the computational model of emotion developed in this thesis should be viewed as a common foundation upon which more sophisticated models of emotion can be developed. The model should not be considered as being a definitive computational model of emotion since there are aspects of emotion elicitation and effect not considered (as mentioned earlier in this section).

The emotions modelled are defined as being composed of six elements: eliciting condition, potential, activation threshold, saturation, effect and probability of effect. These elements along with others that have been identified in other works but not implemented in this thesis are the subject of sections 5.3.1 to 5.3.4.

5.3.1 Emotion Elicitation

As explained in the introduction to this section, the emotion elicitation function implemented is intended to be simple. Since I have taken inspiration from Reilly's computational model of emotion [155] an agent does not undertake a complex reasoning procedure where the consequences of an event or action is repeatedly appraised against an agent's goals or standards. Therefore, the appraisal procedure prescribed by the OCC model [142] is simplified so that the appraisal of the consequences of an event or action with respect to the agent's goals or standards has essentially been performed. Instead, an agent simply has to recognise the event perceived for an emotion to be elicited. To clarify, the appraisal performed by an agent is qualitative and proceeds as such: "I will increase my emotion potential by an amount, x, since my opponent has performed an action α" rather than: "I will increase my emotion potential by some amount, x, since my opponent y has performed an action, α, which I approve/disapprove of since it upholds/violates my standard, S, and its consequences are desirable/undesirable with respect to my goal, G". The appraisal procedure used in this thesis is possible since

the Prisoner's Dilemma game contains very few events and actions and there have been numerous pieces of psychological research undertaken to establish what emotions are experienced in this game given certain events or actions (there has already been some discussion of such work in sections 2.4 and 2.5 of chapter 2).

Therefore, although the emotional model proposed is based upon the OCC model [142], the agents in my simulations do not maintain explicit representations of distinct goals, standards and attitudes. This may seem remiss given that these concepts are proposed as the driving force behind emotion elicitation in [142], but, as highlighted in the previous example, whilst these concepts are not explicitly represented they are implicit in the eliciting conditions of the emotions implemented. This is why the general principles of the emotional model proposed can be directly reused by others but particular aspects can not. Particular eliciting conditions for particular emotions will be discussed fully in subsequent chapters.

Each emotion implemented shares the same general eliciting condition i.e. an emotion is elicited when the value of its potential surpasses the value of its activation threshold (emotion potentials and activation potentials are discussed in further detail in section 5.3.2 below). Thus, I have chosen to model the emotion elicitation mechanism quantitatively primarily because this is the method prescribed by Ortony at al. in [142]. Furthermore, researchers that utilise the OCC model as a basis for their own computational models of emotion also construct their emotion elicitation mechanisms so that they are quantitative. For examples of this, see the discussions in chapter 4 of Steunebrink et al.'s logical formalism in section 4.2.1 along with the computational emotional models of Reilly, Velásquez, Bazzan and Bordini, and Burghouts et al. (sections 4.3.1, 4.3.3, 4.3.7 and 4.3.8 respectively). The advantages of modelling emotion elicitation functions quantitatively rather than qualitatively will be discussed in section 5.3.2 so no further attention will be paid to this issue here.

With respect to all emotions modelled, each event or action that affects an emotion classes' potential is distinct i.e. the event/action that elicits anger in an agent does not elicit gratitude, admiration or hope. Also, each emotion class only has a single eliciting condition, unlike Bazzan and Bordini's emotion elicitation mechanism discussed in [12] and section 4.3.7 of chapter 4. In [12], the eliciting conditions for distress in an agent (A1) are twofold:

> "we test if A1 has not collected at least y = 9 points in the last [previous] round, or at least x = 3 neighbours are experiencing distress. In the affirmative case, A1 might experience distress too"[1]."

My decision to implement emotion classes in this way was again due to a consideration of simplicity since multiple eliciting conditions for the same emotion class can cause significant confusion for human validation of the computational emotional model. By

[1] cf. "*Emotion contagion*" as explored in [124] by Masthoff and Gatt.

ensuring that an emotion class' elicitors are singular and unique it can be ensured that an emotion has been elicited for one reason in particular.

At this point it should be noted that emotion instances are modelled in a way analogous to Reilly's demons [155] or Velásquez's proto-specialists [205] (discussed in sections 4.3.1 and 4.3.3 of chapter 4 respectively). This is primarily due to the elicitation mechanism by which an emotion instance evaluates whether or not it should become active. Since the Prisoner's Dilemma game is limited with respect to the possible events/actions that may occur and, since the implementation details for the emotions modelled are based upon psychological research, each of the agent's emotion instances is represented as a numerical counter (a form of memory). The value of a counter increases or decreases according to the events that occur whilst the simulation is running. In this way, the value of a counter acts as an emotion instance's potential.

My decision to implement emotion instances in this way was taken since Netlogo forbids the implementation of agents inside agents therefore, implementing faithful recreations of Velásquez's proto-specialists [205] is impossible. The closest method that may be used to implement the emotions modelled is to create a specific function in the simulation code that checks the events or actions of others in the round just played and alters relevant emotional instance counters accordingly. A future improvement to this work would be to create a form of listener that updates relevant counters as events or actions occur (like Reilly's demons [155]) without the function being called at explicit points. However, in the simple round-based context of the Prisoner's Dilemma where events and actions occur at fixed points, what I have implemented is sufficient.

It should also be noted that questions regarding which emotion is being displayed in certain circumstances should be asked in context of whether the replacement emotion exists in the OCC model (since this is the psychological model of emotion underpinning the emotion model used this thesis). When deciding upon emotions to model, I take into account experimental evidence from various pieces of psychological studies that analyse and determine the eliciting conditions and effects of various emotions elicited in human-beings when engaged in public-goods games. This evidence is used to determine which of the OCC emotions it is appropriate to model.

5.3.2 Potential, Activation Thresholds and Saturation

For the terms "potential" and "activation threshold" I adopt the definitions given by Ortony et al. in [142]. Since emotion modelling is quantitative with respect to [142], an emotion instance's potential is a variable numerical value associated with an emotion (see section 5.3.1 above). An emotion's activation threshold on the other hand is a constant numerical value (different emotional characters prescribe different constants, see section 5.4).

Quantitative modelling of emotion elicitation mechanisms as opposed to qualitative modelling is advantageous for a number of reasons: firstly, such an approach allows for fine-grained control of emotions. By modelling emotion elicitation quantitatively,

distinction between degrees of emotion potentials is possible. Qualitative modelling on the other hand would appear to assume (based upon the research discussed in section 4.3 of chapter 4) that, if an event occurs and is an eliciting condition for *emo*1, then *emo*1 will *always* be activated. Secondly, quantitative modelling also allows for accurate testing of the model's implementation (in terms of computational implementation rather than its adherence to reality) and for the implementation of emotional characters (see section 5.4).

Along with potentials and activation thresholds, the emotions modelled in this thesis are also endowed with constant numeric "saturation" values. The concept of saturation values is most explicitly considered and implemented in Velásquez's and Burghouts et al.'s computational emotional models (see sections 4.3.3 and 4.3.8 of chapter 4 respectively). An emotion's saturation value prevents its associated potential from exceeding a certain value; by implementing such a concept, emotions cannot become infinitely intense. If this were allowed then significant problems may emerge, for example: an agent could potentially perform the intentional behaviour associated with an emotion infinitely. Anecdotal evidence can be used to show that people have some form of saturation point with regards to their emotions since people do not become infinitely angry; a point is usually reached where the emotion peaks and then begins to decay. In the computational model of emotion proposed in this thesis, an emotion's saturation value is equal to the value of its activation threshold. This decision was made for two reasons: firstly, I do not implement any sort of function for emotional decay (see section 5.3.4 for a discussion of this). Secondly, the psychological research discussed in section 2.4 of chapter 2 shows that: the more intense the emotion, the more intense the action associated with it. In the Prisoner's Dilemma however, a player may only cooperate or defect, so it is impossible to vary the intensity of cooperation or defection exhibited. Therefore, considering emotion intensity in this case would be redundant. Consequently, the agents implemented in this thesis may not experience emotions with different intensities since there is never any difference between an emotion's current potential and its activation threshold when the two values are equal[2].

Lack of consideration for emotional intensity results in the intensity variables identified by Ortony et al. in [142] not being implemented in this thesis. This decision is further justified by my desire to keep the emotional model simple for the reasons already discussed and the criticism noted by Reilly in [155] towards Elliott and Siegle's work on intensity variables in [49]. Since intensity variables and associated functions have not been adequately researched and formalised by those in the psychological field, I think it wise to not include any concept of them here. This does not imply that I consider intensity variables to be irrelevant or their existence to be questionable, rather, I do not consider them necessary for my purposes or sufficiently understood to warrant their inclusion.

[2]Note that emotion *intensity* and emotion *potential* are distinct concepts. The OCC model [142] defines emotion intensity as the degree to which an agent's emotion potential has surpassed the emotion's threshold (see section 4.1.1.1 of chapter 4).

Related to the issue of intensity variables is the function underlying how emotion potentials are calculated in this thesis. When an emotion's potential is increased, it is increased by a constant amount (constant both with respect to a single emotion class and across all other emotion classes) according to the actions performed by other agents i.e. "Agent y defected/cooperated so my relevant emotion instance potential will increase by some value x". Therefore, the same action should consistently alter the emotion's potential since there is no effect exerted upon potential alteration by context. With regards to decreasing an emotion's potential, the issue is a little more complicated due to the effects of opposing emotions. As such, the function for decreasing each emotion class' potential is discussed in chapters 6, 7 and 8.

5.3.3 Emotional Effect and Probability

Following the discussions in chapters 2, 3 and 4, I model the effect of an emotion as being focused solely upon the intentional behaviour of the agent that the emotion has been elicited in. Due to this design decision, the emotions modelled in this thesis may be considered as functional strategies that are utilised by agents when some eliciting condition is met. Answers to the questions of how and why emotions affect the associated intentional behaviour in the way proposed will again be postponed until subsequent chapters. At this point however I will clarify some important aspects of the emotional effect mechanism implemented.

The fact that an emotion's effect is focused upon the intentional behaviour of an agent serves to highlight the functional view I take with respect to them. As has already been mentioned in section 5.3.1, the emotion-behaviour mapping in this thesis is modelled as being one-to-one i.e. activation of anger simply causes the agent to defect, it does not also cause the agent to perform another action. This design decision does not seek to imply that this is the correct or only way to implement emotions, it simply means that, after due consideration of relevant research, this is a suitable way to model the effect of this emotion in context of the simulation proposed. Furthermore, whilst an emotion's effect may be general to the emotion class (anger causes defection, for example), activation of one instance of anger does not entail activation of all instances of anger in an agent. Therefore, the effect of an emotion is limited to one opponent due to the way in which emotional instances are implemented (see the introduction to section 5.3)[3].

Whilst the effect of an emotion class is primarily focused upon the agent's intentional behaviour it may also be focused upon another emotion. In such a case, the effect of activating an emotion instance towards an opponent is to deactivate the currently active opposing emotion instance towards the same opponent, an idea most explicitly implemented by Velásquez (see [205] and section 4.3.3 of chapter 4). Emotions capable of exerting this effect are outlined in chapters 6 and 8 therefore, no further attention will be given to this issue here.

[3]cf. "*Emotional contagion*" [124]. This phenomenon is not explored in the thesis but might well be a fruitful topic for future exploration.

It is also important to note that the effects of emotions implemented in this thesis have a *chance* of occurring and whilst the effects of some emotions are guaranteed, others are not. The decision about whether to make a particular emotion's effect guaranteed or probabilistic is informed by both psychological research and previous implementations of computational emotion models by others (see Frijda's [67], Scherer [167], Bandura [8] and Zoumpoulaki et al. [216]). Again, the specifics of this issue with respect to the emotions implemented are discussed in chapters 6, 7 and particularly in chapter 8.

5.3.4 Emotional Decay

I now turn attention to the issue of emotion decay. Decay functions that have been implemented in the computational models of emotion discussed in section 4.3 of chapter 4 are somewhat pragmatic (see sections 4.3.1 and 4.3.3 respectively). In a similar vein, Ortony et al. outline a simple mechanism in [142] that is intended to act as a function for emotional decay. This function is outlined below and has been slightly edited so as to be more understandable:

> IF emotion-potential > emotion-threshold
> THEN emotion-potential = emotion-potential emotion-threshold
> ELSE emotion-potential = 0

According to this function therefore, not all emotions decay at the same rate since some may have different thresholds. This would support Reilly's comment in [155] that:

> "...the decay process might be quite different from emotion to emotion. Some emotion structures might decay quite slowly, such as anger in characters known for holding grudges. Some might decay quite quickly, such as startle."

Essentially, if an emotion has been activated suddenly it may decay at a quicker rate than it would if it was activated slowly over a period of time. Establishing the specific decay rates for particular emotions in particular contexts has not yet been attempted by psychologists (to the best of my knowledge).

Masthoff and Gatt in [124] consider the effect of emotion decay in the context of designing functions to model affective states for users in group recommender systems. It is noted that decay (δ) is likely to depend on personality and three variants of the affective state function are proposed. Of these functions, the authors discover that the third variant performs best and they confirm that rates of decay depend on the individual. Masthoff and Gatt also note that limited task duration reduces the value of δ therefore, time (in this case, the time in which an emotion eliciting experience is present) has a direct impact upon the rate of emotional decay.

From the discussion thus far it would appear that decay functions in current implementations are entirely pragmatic:

- Velásquez and Reilly propose reducing an emotion's potential by one every tick.

- Steunebrink et al. propose using a sigmoid function.
- Ortony et al. propose the function described above.
- Masthoff and Gatt propose three variants of a function to model affective states that include decay and find that the third variant fits their particular experimental data best.

Given the information discussed thus far, it seems likely that there may not be any function for emotional decay that is objectively correct given the array of functions proposed. In addition, Masthoff and Gatt's proposal that the rate of decay for an emotion is likely to differ from person to person adds further complications. As can be seen in the relevant sections of chapters 6, 7 and 8 where particular emotions are discussed; no research mentioned even takes into account possible upper and lower bounds of values for emotional decay. Therefore, it is not possible to even specify a range of experimentally-informed values that emotion decay can be set to.

Whilst these points regarding the subjective nature of emotion decay inform my decision on whether or not to model emotional decay in this thesis, the main consideration regards *time*. As pointed out by Masthoff and Gatt in [124], the length of time that an emotion eliciting event lasts for plays a crucial role in determining the extent to which emotional decay affects the current potential of emotions. In section 5.3.3 it is made clear that the events which affect emotions in the test-bed used in this thesis last for a very short period of time and occur so frequently that an agent's emotions are *constantly* updated. Therefore, implementing a function to cater for emotional decay is unnecessary in such circumstances since emotional decay would not have a chance to occur. Decay is important, but only when the eliciting conditions occur infrequently over a relatively long period of time, so that decay has an opportunity to take effect.

Not implementing a decay function results in different play histories producing the same emotional effect. For example, if an agent's activation threshold for anger is set to 3 and its opponent defects three times and cooperates twice, then anger is activated in that agent no matter what order these cooperations/defections occur.

5.4 Emotional Characters

Emotional character is a concept that has arisen several times in section 4.3 yet the term "emotional character" is not always used. It is my intention here to briefly discuss the definitions presented by others before moving on to present my own definitions of "character" and the associated term "characteristic".

Perhaps the most explicit example of an emotional character implemented in the literature discussed in section 4.3 of chapter 4 is the one presented by Poel et al. and Burghouts et al. in [149] and [25] respectively. In [149], Poel et al. introduce the term "character" which specifies the emotions an agent can experience, how strongly/weakly an agent experiences emotions and how long an agent's emotions are experienced for.

An agent's, "character" also refers to the set of typical behavioural strategies it can apply in particular situations.

In [25], Burghouts et al. appear to drop the concept of "character" and instead focus upon "personality". In defining "personality", Burghouts et al. state that they take inspiration from Lisetti and Gmytrasiewicz [74]. Consequently an agent's personality defines the agent's "attitude" with respect to its optimism, confidence and extraversion.

In section 4.3.3 of chapter 4, it was shown how Velásquez gives users the ability to alter each proto-specialist's activation threshold value, saturation value and decay function in the Simón simulation. Manipulation of these proto-specialist values enables users to create individual "affective styles" resulting in different emotional reactions given the same circumstances. This definition appears to incorporate elements of the emotional experience in much the same way as Poel et al.'s definition of "character" in [149] does and is interesting since the concepts discussed are not as abstract when compared to Burghouts et al.'s definition of personality. For example: the "appraisal" of events and the "optimism", "confidence" and "extraversion" of an agent are either qualitative in nature or are not described in detail in the papers that mention them. The elements that constitute an agent's "character" however are much more concrete and well-defined as they are quantitative in nature and have agreed upon implementation methods.

It is also important to note that characters, personalities and affective styles all impart some degree of individuality for an agent by allowing variability of elements that constitute an emotional experience. This serves to reflect that emotional reactions are rarely consistent in any given population and it is therefore important to suggest a concept in this work by which distinct emotional characters/personalities can be modelled, implemented and analysed in the simulations proposed. I therefore propose to define emotional character in terms of *activation thresholds*.

5.4.1 Emotional "Characteristic"

The "characteristic" of an emotion is intended to translate an emotion's numerical activation threshold value into a lexical form. By doing this, I hope to be able to facilitate both the description and discussion of an agent's emotional state. There are therefore two constituents which, when combined, comprise the description of an emotion's characteristic. These constituents and how they combine to form a characteristic are outlined below.

Since an emotion's characteristic is intended to lexically represent the value at which an emotion's activation threshold is set, the first part of an emotion's characteristic is based upon the numerical value of an emotion's activation threshold. Regardless of the emotion, activation thresholds may only ever be set to the values of either 1, 2 or 3 and each of these activation thresholds is paired with an associated descriptive noun, either "less", "moderately" or "highly". Consequently, it is this descriptive noun that forms the first element of an emotion's characteristic. The decision to use the values 1, 2 and

3 is completely arbitrary; there exists no definitive research to prescribe what values should be used in this case. At this point it should be noted that the descriptive noun used is emotion-dependent hence the reason why activation thresholds of 1 or 3 may be set to either "less" or "highly". The activation threshold values and their associated descriptive noun are listed below:

- 1: "less/highly"
- 2: "moderately"
- 3: "less/highly"

The second part of an emotion's characteristic is an adjective that is associated with the emotion class. These adjectives are not based upon any previous work by other researchers; they are novel. The emotion classes implemented and their paired adjectives are listed below.

- Anger: "tolerant"
- Gratitude: "responsive"
- Admiration: "impressionable"
- Hope: "greedy"

The best way to illustrate the formation and use of a characteristic is with an example: if an agent's activation threshold for anger is set to 1 then the agent's emotional characteristic for anger can be defined as "less tolerant". If the activation threshold for anger is set to 2 then the agent's anger characteristic can be said to be "moderately tolerant". The characteristic of an agent whose activation threshold for anger is set to 3 is "highly tolerant". Therefore, whilst the adjective remains constant for an emotion class, the adverb changes depending upon the value of that emotion's activation threshold. Use of these emotional characteristics enables explanations of agent behaviour to be expressed in a more natural way.

Ortony et al. note in [142] that the alteration of activation threshold values allows observations such as "John was in a wonderful mood that morning. When his children were obnoxious at breakfast, it didn't bother him at all" to be made sense of since it may be assumed that obnoxious children would anger John. However, if John's good mood that morning causes his activation threshold for anger to rise then it makes sense to say that John is more "tolerant" than usual. Likewise, if John was quick to anger then it could be said that John is less tolerant than usual yet as always, such comparisons are relative to a norm. In the context of the simulation it is assumed that having an activation threshold value of 2 is the norm and all other activation thresholds are defined in context of this. This value was not chosen based upon any previous psychological work but simply because it is the central value of the three possible activation threshold

values. The requirement was simply to enable three speeds of reaction, so that we could see agents playing with agents who responded more quickly/as quickly/more slowly with respect to both anger and gratitude, and the numbers chosen offer a pragmatic way to achieve this. Similarly, the activation thresholds chosen are completely arbitrary since I know of no attempt made to associate real numbers with activation thresholds in the context of human emotion. They do, however, correspond with some attitudes found in common parlance: "zero tolerance" for an immediate reaction; the proverb "fool me once, shame on you; fool me twice, shame on me"; "three strikes and you are out" which has migrated from baseball to legal use.

5.4.2 Emotional "Character"

The "character" of an agent is intended to allow the precise and succinct description of all the emotional characteristics associated with an agent. There are nine basic characters implemented in this thesis extending up a maximum of twenty-seven with each character being denoted alphanumerically. The characters implemented in this thesis and their specifics will be discussed in subsequent chapters however, the nine basic characters are defined in terms of the anger/gratitude characteristics that an agent possesses. To aid visualisation and conceptualisation of these emotional characters I will make use of a 2-D matrix in chapter 6.

To give a brief example of how an emotional character is defined consider an agent whose activation threshold for both anger and gratitude is set to 1. In this case the agent would have two characteristics defined; on the one hand its characteristic for anger can be described as "less tolerant" whilst its characteristic for gratitude can be described as "highly responsive". Constantly repeating these characteristic definitions becomes laborious and confusing, especially when discussing the behavioural patterns demonstrated by agents with different emotion activation threshold configurations. Therefore, the agent described above may be referred to as A1:G1[4] where A1 indicates that an agent's threshold is set to 1 and G1 indicates that an agent's gratitude threshold is set to 1.

5.5 Chapter Summary

The aim of this chapter was to provide an overview of the research questions, simulation test-bed and emotional model that will be answered/used in the remainder of the thesis. This chapter also serves to make explicit the novel aspects of my work compared with the work of previous researchers who have performed similar investigations. As stated in section 5.1, my general research interests lie in the study of how emotions may affect interactions amongst social agents.

In section 5.2 I also attempted to give a general discussion of the simulation test-bed that will be used to gather data that is capable of answering the research questions posed

[4] **A** and **G** are contractions of **A**nger and **G**ratitude respectively.

in section 5.1. As mentioned, the decision to create and utilise a simulation test-bed was made since simulation appears to be the *de facto* choice for researchers who are looking to answer research questions using emotional agents. Simulation provides a number of benefits including precise control of simulation variables so that there are no *uncontrolled* variables (or at least very few that exist in the simulation itself) and efficient methods to facilitate the gathering of large sets of data from repeated experiments (when compared to most other methods of investigation). The simulation test-bed developed is based upon the iterated version of the Prisoner's Dilemma game which provides the flexibility to explore a range of issues. For example: interaction with multiple adversaries is required for the modelling of admiration in chapter 7.

One of the main contributions of this thesis is the use of the emotional modelling framework and the concept of emotional character presented in sections 5.3 and 5.4 to drive a coherent series of simulations. These simulations will explore the effects of several emotions on behaviour of agents from both societal and individual perspectives, as will be described in chapters 6, 7 and 8. The emotional model is predominantly based upon the OCC model of emotion [142] which has been used by a number of other researchers interested in the computational modelling of emotion. Whilst this is not revolutionary, some details of the emotional model described in section 5.3 have originality when compared with other computational implementations of the OCC model. For example: agents do not have one emotion instance for all opponents, instead agents have an emotion instance for each opponent. Emotion instances are therefore *discrete* and *directed* facilitating rich and flexible intentional behaviour from an agent towards its opponents. Furthermore, I have presented the six basic elements of an emotion that must be computationally modelled in order for an emotion to be simulated in section 5.3, these are: eliciting condition, potential, activation threshold, saturation, effect and probability of effect.

In section 5.4 I have defined emotional character in terms of activation thresholds *only*. An agent therefore has a number of emotional characteristics equal to the number of emotion classes it is endowed with. An emotion's characteristic acts to lexically describe the value at which the emotion classes' activation threshold is set to. An emotional character on the other hand enables a succinct summation of an agent's emotional characteristics by using an alphanumeric identifier. By defining agents in terms of their emotional character I am able to easily reference and discuss particular emotional features of an agent in a succinct and understandable manner.

The key points to be taken from this chapter are listed below:

- The work in this thesis can be classed as being part of the *cognitive-engineering* and *experimental-theoretical* research motives defined by Burghouts et al. in [25]. My work is *not* part of the believable-agent research motive since I am not interested in suspending belief in humans that the agents constructed are machines.
- To investigate the research questions posed I will make use of a *simulation* test-bed based upon the *iterated Prisoner's Dilemma* game.

- The computational model of emotion developed is based predominantly upon the *OCC model* of emotion [142].
- Any emotion can be modelled by specifying its *eliciting condition*, *potential*, *activation threshold*, *saturation*, *effect* and *probability of effect*.
- Appraisals of events that may give rise to emotion are *qualitative* whilst the emotion elicitation function itself is *quantitative*.
- The appraisal process is *simplified* so that there is no explicit appraisal of how the consequences of an event or action affects an agent's goals or standards. These appraisals are *predefined* and *implicit* given the limitations of the Prisoner's Dilemma game.
- When activated, emotions have a *probability* of directly affecting the agent's current *intentional* behaviour.
- Emotional decay is *not* modelled because the events that elicit emotion occur close together in an uninterrupted sequence and so arguably, there is no period of reflection in which decay can take effect in the current set-up.

Chapter 6

Anger and Gratitude

This chapter is the first of three that consider each of the individual simulations run using agents endowed with emotions developed using the emotional model presented in chapter 5, section 5.3. These three chapters consider the effects of different emotions and emotional characters upon the *success* of those agents in different scenarios. It should be noted that the term "success" takes a number of different meanings throughout the remainder of this thesis and one should not ascribe any universally applicable definition to it. In view of this, "success" will be defined appropriately for each simulation context.

The work presented in this chapter aims to make three main contributions: the first pertains to the methodology of modelling emotions using the computational model of emotion outlined in chapter 5, section 5.3 as a basis. By using a combination of the emotion model proposed and existing psychological research, I show how emotions are able to play a functional and beneficial role (both from the perspective of individuals and the total system) in enabling agents to respond to information received from the environment. Secondly, I intend to demonstrate that emotions are an important contributing factor in enabling the phenomenon of cooperation to occur in MAS and that the emergent property of cooperation can be achieved through emotional response alone, without consideration of past, present or future pay-offs. Finally, I identify and discuss the characteristic behaviour of various emotional characters to provide a full examination and understanding of how emotions can influence the behaviour of individuals engaged in a social dilemma context.

To realise these contributions I use Axelrod's computer tournament [7] as inspiration (discussed in section 2.4.3 of chapter 2). To quickly recap, Axelrod uses the iterated version of the Prisoner's Dilemma game to test the effectiveness of various strategies to enable cooperation. The tournament and its subsequent discussion is a foundational piece of work concerning cooperation and how it may evolve in societies of purely non-emotional, self-interested individuals. This work therefore serves as a suitable base upon which my own investigation may be built since it shows that, whilst self-interested behaviour can enable cooperation, researchers such as Frank, [62], argue that such behaviour can be self-defeating. As already discussed at length in chapter 2, individuals endowed with emotions are much more likely to establish and maintain cooperation but

the question of what emotions are involved is debatable, as is the question of why these emotions are important.

This chapter is divided into six sections: section 6.1 outlines the research questions that the investigation presented in this chapter will attempt to answer. Section 6.2 discusses the emotions that are implemented in this chapter along with justifications for their inclusion and how they are computationally modelled using the emotion model presented in chapter 5, section 5.3. Section 6.3 presents the emotional characters implemented and used in this chapter whilst section 6.4 provides information about the simulations used. Section 6.5 contains an in-depth analysis of the results obtained by the simulations, a discussion of how these results emerged and how they relate to the research questions listed in section 6.1. Finally, section 6.6 summarises the main conclusions gleaned from this chapter.

6.1 Research Questions

1. What emotions should be modelled to allow agents to respond to the actions of others in Prisoner's Dilemma games?

2. How should these emotions be modelled computationally using the emotion model proposed in chapter 5, section 5.3?

3. How should the base emotional characters that are to be used in the rest of this thesis be modelled?

4. Do any of the emotional characters modelled offer an improvement upon the success exhibited by the TFT strategy, which was identified as the most successful strategy in Axelrod's tournament? If so, why?

5. Are there any other interesting behavioural features associated with the base emotional characters proposed in this chapter? If so, what are they and why do they exist?

6.2 Emotions Modelled

If I am to create accurate models of emotion, I must consider each emotion that I have chosen to implement carefully in order to correctly ascertain their eliciting conditions, effects and how to alter their potentials in the context of public goods games. The amount of literature that considers emotional responses in Prisoner's Dilemma games is quite sparse so my discussion of emotions in public goods games will not just be limited to the Prisoner's Dilemma. My intention is instead to consider a broad array of public goods games so that I can provide sufficient evidence to assert that the emotions I have chosen are appropriate and details of their implementation are valid. By the conclusion of this section, the work undertaken should have provided answers to research questions

1 and 2 from section 6.1 i.e. *"what emotions should be modelled to allow agents to respond to the actions of others in Prisoner's Dilemma games?"* and *"how should these emotions be modelled using the emotion model proposed in chapter 5, section 5.3"*.

The algorithm which controls the activation of anger and gratitude is specified in detail in section 6.4.4. This is because additional information regarding the simulation set-up is required to fully understand its operation. Furthermore, there is no discussion in the following section about results obtained by others who have modelled anger and gratitude in a MAS context. This is because there simply is no research to comment upon. As outlined in section 5.1 of chapter 5, the four relevant pieces of research [12, 25, 95, 139] either do not have results that relate to the context of the simulations used in this thesis [139] or do not give any details as to the effects of anger and gratitude [12, 25, 95].

6.2.1 Anger

Research that considers the emotion of anger in the context of public goods games is quite definitive in its observations and conclusions. Despite this, it is difficult to draw a clear line of separation between research that considers the eliciting conditions of anger and those that study its effects. Therefore, I will discuss both aspects simultaneously in section 6.2.1.1 along with research that supports my decision to model this emotion. Section 6.2.1.2 will then describe how the salient points outlined in section 6.2.1.1 will be used to model the emotion computationally using the emotion model proposed in chapter 5, section 5.3 as a framework.

6.2.1.1 Eliciting Conditions and Effects

According to the OCC model of emotion [142], anger is elicited when an agent disapproves of another's blameworthy action and is displeased about the related undesirable event. However, determining the eliciting conditions for anger in the context of public goods games would appear to be difficult to specify. Since such issues are highly subjective it is necessary to consider all viable options before moving on to specify how I will objectively model the emotion in a computational context. I begin by considering Fehr and Gächter's research [57, 58] which has been mentioned in sections 2.4.1 and 2.4.1.1 of chapter 2. I will focus upon [58] as this work is a condensed version of [57].

In [58], Fehr and Gächter attempt to determine if emotions are the proximate cause of the behaviour observed in their experiments by asking participants to consider a hypothetical situation and indicate their emotional response to it. There are some issues that I take with such a method of ascertaining emotional states, these are outlined by Tsang in [201]. Firstly, hypothetical situations may have low psychological realism causing participants to not respond realistically to the events that occur within them. Secondly, these emotional responses do not have any real cost to the participant therefore, factors such as "judgement from others in response to an emotional experience" do not have any bearing upon the emotions reported. Thirdly, participants may not even experience

the emotion reported; instead, they may anticipate that they would and then record the anticipated emotion and its anticipated potential.

Irrespective of these drawbacks, the results from Fehr and Gächter's emotional questionnaire propose that the proximate cause of anger in public goods games is the recognition of inequality between players. The issue with this conclusion is that the emotions reported do not take into account the individual *actions* of opponents since these are not stated in the hypothetical scenario presented. Instead, the only information the responder has to inform their emotional response is the pay-off they are presented with. Indeed, it would appear that a significant amount of research which proposes that inequality evokes anger suffers from this flaw. For example: Dawes et al. conduct a series of experiments to investigate whether inequality is the salient motivation for the occurrence of anger in [37] and conclude that:

> "...individuals apparently feel negative emotions towards high earners and the intensity of these emotions increases with income inequality."

This conclusion is problematic since the emotions of the participants are not recorded after every round of play instead, participants are presented with hypothetical scenarios practically identical to those used by Fehr and Gächter *after* the game. Therefore, the conclusion proposed is based upon a consideration of the emotional effect incurred by the *pay-offs* received by others. No consideration is afforded to how the *actions* of an opponent may affect the emotions of a participant and their subsequent actions.

It would appear that there are few pieces of research that specifically consider the effects that individual actions may have upon participants in public goods games. To date, I have only unearthed one piece of research that addresses the issue directly: Burton-Chellew et al.'s paper, [26]. Whilst the experimental set-up implemented is very similar to that used by Fehr and Gächter in [57] and [58], participants were instead asked to report their emotional state with respect to anger and guilt after every round of a *real* public goods game being played. Burton-Chellew et al.'s observation that individuals felt more angry when they contributed relatively more than their group mates would indicate that free-riding or *defection* is capable of eliciting anger in the course of a public goods game. This is not to say that inequality does not evoke anger in participants, but defection would appear to be a more fundamental eliciting condition in this context.

Despite these differences, all research discussed so far appears to agree with the effects of anger, namely that it motivates some form of punishment. Sanfey et al., [164], observe that participants with stronger anterior insula activation (an area of the brain associated with negative emotional states; particularly anger and disgust) to unfair offers in Ultimatum games reject a higher proportion of these offers whilst Dawes et al. in [37] note that participants who indicate feelings of anger spend 26% more to penalise above-average earners than subjects who said they were not annoyed or angry. A further study on the eliciting conditions and effects of anger is undertaken by Small and Loewenstein who examine the effects of identifiability (making known the identity of a person) upon the intensity of anger elicited in public goods game participants [178].

Small and Loewenstein use a similar public goods game set-up to those used by Fehr and Gächter in [57] and [58]. Important points to note about the experimental set-up are:

- Punishment may only be exacted by group investors upon non-group investors.
- Punishment is costly to the punisher.
- Immediately after punishment, punishers were asked to indicate how intense their emotions of anger, blame and sympathy were with respect to the non-group investors they punished.

The hypothesis that contributors would apply harsher penalties in the identifiable condition (where the identity of non-group investors is known) than in the unidentifiable condition (where the identity of non-group investors is not known) is supported by the results obtained by Small and Lowenstein. Whilst the link between identifiability and intensity of anger experienced is of no relevance to my simulations, it is interesting to note the relationship between anger intensity and the degree of punishment exacted. With respect to this, participants reported more intense anger in the identifiable condition than in the unidentifiable condition. Ergo, it would appear that anger and punishment have a direct relationship to each other: the greater the intensity of anger experienced towards a person the harsher the punishment exacted towards them. This relationship is noted by Soloman in [184] where it is also stated that anger naturally induces a desire to punish and by Pilluta and Murnighan in [148] where it is demonstrated how anger motivates the punishment of proposers in Ultimatum games who make offers that are deemed unsatisfactory by the responder.

The observation that anger provokes punishment in humans also appears to exist across different cultures with similar results and conclusions obtained by Shinada et al. [172] in Japan. Once again, Shinada et al.'s experimental set-up closely mirrors Fehr and Gächter's [57], [58]. In [172], participants play a version of the Dictator game where punishment is costly and punishers are asked to report their feelings of anger towards the participants punished following the game's conclusion. The paper states that there is a positive correlation between the anger experienced towards free riders and the degree of punishment exacted given the results obtained.

As a final comment, it is worth pointing out one of the most all-encompassing works available on the subject of anger and punishment in public goods games. Sigmund's review of the topic, [173], gathers together evidence from a number of different public goods games where punishment is possible and is, at times, costly. Two conclusions from this work are particularly relevant to the current discussion:

- Being cheated or being treated unfairly in public goods games arouses negative emotions such as anger in human players.

- Negative emotions cause costly punishment to be exacted upon those who attempted to secure an advantage over the player or those who benefited in an unfair split of reward.

6.2.1.2 Modelling Anger Computationally

The research considered in section 6.2.1.1 provides evidence that defection from opponents in public goods games is capable of eliciting anger whilst anger can cause an agent to punish the opponent responsible for its elicitation. Furthermore, the research outlined in section 6.2.1.1 asserts that anger causes human players to punish opponents in public goods games even when such punishment is costly.

If two agents, agent x and agent y, are opponents in a Prisoner's Dilemma game then, if y defects, x's score will suffer. If y defects then a cooperating x will receive 0 rather than 3, and a defecting x will receive 1 rather than 5. Considering this in conjunction with the results of research discussed in section 6.2.1.1, I have modelled *defection* from an opponent as being the eliciting condition of anger for an agent in the simulations.

The example above would then seem to fit with the eliciting condition for anger outlined in the OCC model (see the introduction to section 6.2.1.1): agent x disapproves of y's defection as it is selfish and is displeased since this defection reduces x's pay-off, whether x cooperates or defects. As was mentioned in section 5.3 of chapter 5, the determination that y's defection is selfish and that x disapproves of this defection since it entails that x can no longer achieve its most preferred outcome, is implicit. Such a design choice was made due to the research discussed in section 6.2.1.1: those who witness defection appear to consistently disapprove of this action because it violates some standard: "contribute a similar amount of money as others" or "do not be selfish", for example. It is worth mentioning that, even if x has also defected in a round when y defects, x's anger potential towards y will still be increased by the same amount as it is even if x were to have cooperated. This design choice is a notable limitation and is a result of attempting to limit the complexity of the model in these early stages.

As alluded to in the above example, defection is also a form of punishment in the Prisoner's Dilemma since defecting against an opponent can only ever result in the opponent achieving one of its two least preferred outcomes: DD or DC. Whilst defection may be profitable, it may also be very costly and is risky, especially in the long term if it leads to a lengthy sequence of DD. In the long run avoiding convergence on DD is the most important aim if a high score is to be achieved against a range of characters. Taking this into consideration I propose that the effect of anger is to cause an agent to *defect*.

With respect to an agent's anger emotion potential, its potential will increase by one every time an opponent defects until it reaches its saturation point. This may seem insufficient given that the majority of the research presented in section 6.2.1.1 argues for anger being experienced more quickly when more intense defection is received. However, given that it is not possible to defect to any greater or lesser degree (by virtue of the

Prisoner's Dilemma game set-up), it seems sensible to implement the effect of defection in this way. In addition to the previous point regarding the absence of degrees of defection by virtue of the Prisoner's Dilemma game set-up, since the emotion modelling framework proposed does not consider emotion intensity for the reasons given in section 5.3.2 of chapter 5 (the temporal proximity of emotion eliciting events are very close), I am not able to implement the differing effects of anger as mentioned in section 6.2.1.1; harsher or more protracted periods of punishment). Providing such refinement is left for future work and as a result this will be discussed in chapter 9.

Finally, based upon the research discussed in section 6.2.1.1, anger is *guaranteed* to cause the agent to defect if its activation threshold is reached and the emotion is elicited. As with the alteration of emotion potential for anger, modelling different probabilities of this emotion's effect is left for future work.

6.2.2 Gratitude

Gratitude and reward, like anger and punishment are inextricably linked in the context of public goods games. As with anger, there have been a number of research papers already discussed in this thesis that provide support for modelling gratitude. Following in the footsteps of section 6.2.1, I first devote attention towards the consideration of research that focuses upon the eliciting conditions and effects of gratitude in section 6.2.2.1. This section will also provide evidence to support my decision to model this emotion in the context of a public goods game. Following this, I will proceed to document details of the computational model for gratitude used in section 6.2.2.2.

6.2.2.1 Eliciting Conditions and Effects

Gratitude, according to the OCC model [142], is elicited by the approval of another agent's praiseworthy action and being pleased about the related desirable event. A number of works have already been discussed in section 2.4.2 of chapter 2 that postulate acts of reward and cooperation as being eliciting conditions for gratitude in others. Heider, [84], illustrates that the intensity of gratitude experienced is greater when an altruistic act directly benefits a recipient rather than witnessing a third-party benefiting from an altruistic act. Tesser, Gatewood and Driver, [196], expand Heider's conclusion by analysing the effects of the recipient's perception of the intentionality of the benefactor, the cost incurred by the benefactor in providing the benefit and the value of this benefit to the recipient on the intensity of gratitude experienced. The results obtained indicate that a person's intensity of gratitude increases as each of the variables listed above are increased and the effects of these variables do not appear to interfere with one another. In [127], McCullough et al. incorporates these proposed eliciting conditions for gratitude by asserting that gratitude is:

> "...a positive emotion that typically flows from the perception that one has benefited from the costly, intentional, voluntary action of another person."

Furthermore, whilst relatively little attention is paid to the consideration of positive emotions by Xiao and Houser in [214] they do note that around 80% of players who took part in their Ultimatum games displayed positive emotions towards proposers who gave fair offers. Unfortunately, there is no identification of the particular emotion in this case so specifying that it is gratitude rather than joy that is elicited by fair offers (for example) is impossible.

This impasse is addressed by the work of Tsang [201] (mentioned in section 6.2.1.1) where gratitude is defined as "*a positive emotional reaction to the receipt of a benefit that is perceived to have resulted from the good intentions of another*". In [201], gratitude is induced in experimental participants and its existence confirmed by way of behavioural measures and self-reports. The emotion is induced to ensure that participants respond emotionally to current, real situations rather than hypothetical situations. The issues associated with using hypothetical situations to gauge emotional responses and propose resulting conclusions were discussed with respect to anger in section 6.2.1.1. Crucially, Tsang's research finds that people are more likely to display generosity towards people who intentionally reward them in some way rather than through chance. Those who were under the impression that their opponent rewarded them intentionally also reported higher intensities of gratitude. The controlled environment used in [201] lends strong support to the notion that gratitude does indeed motivate cooperative pro-social behaviour when it is elicited.

I now turn to consider the effects of gratitude i.e. does gratitude cause intentional cooperative behaviour in the context of public goods games? In comparison to the relatively scant research concerning how gratitude may be elicited in public goods games, the literature that attempts to answer what effects gratitude may have upon players in public goods games is voluminous. Research of this kind has been outlined in chapter 2, with particular reference to sections 2.4.1.2, 2.4.1.3 and 2.4.2. Aside from these works, there are other notable pieces of research that lend their support to the proposition that gratitude motivates players in public goods games to endow opponents with costly rewards. In [79], Guala and Mittone make reference to experimental results obtained in Ultimatum games run by Charness and Rabin in [29]. With respect to [29], these Ultimatum games show how some responders will choose an outcome where they receive less than the proposer ($400 to the responder, $750 to the proposer) rather than an equal amount. Guala and Mittone argue that intuitively, some feeling of gratitude may motivate such a reward for the proposer, since the responder recognises that they could have been offered $0 at no cost to the proposer. Perhaps the most definitive argument produced for gratitude motivating cooperative, pro-social behaviour is provided by McCullough et al. in [127] where a number of papers are cited to support the author's position that gratitude is a motivator of pro-social behaviour. Of note is the discovery by McCullough et al. in [126] that participants who reported higher intensities of grateful dispositions exhibited greater pro-social behaviours. Furthermore, results from Emmons and McCullough [53] show how participants who were asked to write about

events that elicited gratitude offered more emotional support and tangible help to others than those who wrote about events that produced negative emotions.

An analysis of gratitude with respect to both its eliciting conditions and effects in public goods games is provided by Algoe and Haidt in [5] where two out of three studies pertain to the investigation of what these eliciting conditions or effects are. Essentially, their findings fully support the claim that gratitude is elicited by receiving help, kindness or generosity and its effect is to motivate people to acknowledge the source of their gratitude and repay or reward the source directly. Fessler and Haley also discuss these two aspects of gratitude in [60] where research by Berg et al. [18] is cited. Fessler and Haley summarise the conclusions drawn from Berg et al.'s research by stating that individuals respond with gratitude to un-compelled acts of generosity and thus feel subjectively motivated to reciprocate in kind. Furthermore, they indicate how participants in behavioural economics games often violate the assumptions of traditional rational actor models by demonstrating a willingness to incur monetary costs in order to reward partners for perceived cooperative or altruistic behaviour.

In summary, gratitude would appear to be elicited by receiving rewards or cooperation from others that are both intentional and costly to the benefactor in public goods games according to experimental research. Furthermore, gratitude causes individuals to engage in intentionally cooperative and reciprocal behaviour towards others (especially the original benefactor). What is required now is a computational translation of these eliciting conditions and effects in context of a Prisoner's Dilemma game.

6.2.2.2 Modelling Gratitude Computationally

Given the findings and conclusions drawn by the research discussed in section 6.2.2.1 above, the computational modelling of gratitude's eliciting conditions and effects in context of the Prisoner's Dilemma game is relatively simple. Essentially, these eliciting conditions and effects are the opposites of those proposed for anger (see section 6.2.1.2).

With reference to section 6.2.1.2, I will again consider two agents: agent x and agent y, who are opponents in a Prisoner's Dilemma game. So, if x cooperates at any point towards y, y will enjoy a more preferred outcome than if x had defected. To clarify, cooperation from x means that the outcome of the round will only ever result in the two most preferred outcomes for y: $\mathrm{C}_x\mathrm{C}_y$ or $\mathrm{C}_x\mathrm{D}_y$. Furthermore, cooperation from x means that x forfeits its most preferred outcome $\mathrm{D}_x\mathrm{C}_y$. If y also cooperates then this prevents x from receiving its least preferred outcome: $\mathrm{C}_x\mathrm{D}_y$. Cooperation is therefore an intentional, voluntary action performed by an agent that may be potentially costly but rewards the opponent in context of the Prisoner's Dilemma game since cooperation enables the occurrence of another player's two most preferred outcomes: CD or CC. Cooperation's potential cost to the cooperator is embodied in the risk of the cooperator receiving its least preferred outcome DC. Therefore, with respect to the simulations and agents implemented in this thesis, if two agents x and y are considered then, when y *cooperates* with x (for example), x's gratitude potential will increase.

As justified in section 6.2.2.1, gratitude's effect is to endow an opponent with a reward that is potentially costly to the rewarder. As was stated in the previous paragraph, cooperation in the Prisoner's Dilemma may result in the cooperator receiving the sucker's pay-off if its opponent defects. Likewise, if a player cooperates then they will not be able to reap the benefits of distributing a sucker's pay-off to an opponent; either way, cooperation is costly. In context of the Prisoner's Dilemma then, gratitude's effect is to cause the player to *cooperate* when activated.

As with anger, implementing different alterations to gratitude's emotion potential depending upon the joint outcome of a Prisoner's Dilemma round, is not considered. To maintain simplicity I have again decided to increment gratitude's emotion potential by one until its saturation point is reached. Therefore, even though the research presented in section 6.2.2.1 illustrates that the intensity of gratitude experienced increases/decreases in accordance with how costly the reward from an opponent is/how fair the eventual pay-off is, this variation is not implemented. Again, such functionality is left for future work and consequently, this topic will be discussed further in chapter 9.

Finally, with respect to the issue of probability of effect, the elicitation of gratitude is *guaranteed* to cause cooperation in the agent that the emotion is elicited in. This decision is based upon the research discussed in section 6.2.2.1 where the majority of participants in all experiments who experience gratitude respond pro-socially afterwards. Again, as with anger, modelling different probabilities of effect for this emotion is reserved for future work.

6.3 Emotional Characters

Following the definition of anger and gratitude's eliciting conditions, effects, potential calculation and probabilities of effect, the emotional characters implemented in this particular set of simulations can be described. In performing this task I will be supplying an answer to research question 3 from section 6.1, namely: "*how should the base emotional characters that are to be used in the rest of this thesis be modelled?*"

In section 5.4.2 it was mentioned that an agent's emotional character is intended to provide a precise and succinct description of the emotional characteristics associated with an agent. An agent's emotional characteristics on the other hand lexically describe the numerical activation thresholds for the emotions they are endowed with (see section 5.4.1). In this chapter and set of simulations, agents are endowed with two emotion classes: anger and gratitude.

In section 5.4.1 I also explained that an emotion's numerical activation threshold may only be set to either 1, 2 or 3; since each agent is endowed with gratitude and anger in the simulations discussed below, a total of nine emotional characters may be implemented. The emotional characters used are presented in a two-dimensional matrix (see table 6.1) and the emotional characteristics that comprise each of the characters defined in table 6.1 are listed afterwards for reference purposes.

TABLE 6.1: Emotional character definitions.

		If defecting, #coops required to coop.		
		1	**2**	**3**
If coop, #defects required to defect.	**1**	A1:G1	A1:G2	A1:G3
	2	A2:G1	A2:G2	A2:G3
	3	A3:G1	A3:G2	A3:G3

- A1:G1 - Less tolerant and highly responsive (this character is also analogous to Axelrod's TFT agent).
- A1:G2 - Less tolerant and moderately responsive.
- A1:G3 - Less tolerant and less responsive.
- A2:G1 - Moderately tolerant and highly responsive.
- A2:G2 - Moderately tolerant and moderately responsive.
- A2:G3 - Moderately tolerant and less responsive.
- A3:G1 - Highly tolerant and highly responsive.
- A3:G2 - Highly tolerant and moderately responsive.
- A3:G3 - Highly tolerant and less responsive.

Therefore, the intentional behaviour of agents endowed with emotions in all simulations outlined in this thesis is determined by a combination of their *emotional character* and their *current emotional state*. Note that the emotional characters can be divided into two sets based upon their characteristics: the horizontal *tolerance* set and vertical *responsive* set ("horizontal" and "vertical" refers to the placing of the emotional characters that comprise these sets in table 6.1). Each of these sets may then be further sub-divided into three subsets based upon the levels of tolerance or responsiveness of the emotional characters within the sets. Dividing the emotional characters up in this way allows for the isolation and identification of effects that tolerance and responsiveness may have upon individual and total system scores in the discussion of results presented in section 6.5. Tolerance and responsive sets are listed below for clarification and reference purposes:

- Tolerance sets
  - A1:G1, A1:G2, A1:G3 - Less tolerant
  - A2:G1, A2:G2, A2:G3 - Moderately tolerant
  - A3:G1, A3:G2, A3:G3 - Highly tolerant

- Responsive sets
  - A1:G1, A2:G1, A3:G1 - Highly responsive
  - A1:G2, A2:G2, A3:G2 - Moderately responsive
  - A1:G3, A2:G3, A3:G3 - Less responsive

6.4 Simulation Details

The purpose of this section is to describe the main features of the simulations to be run by discussing details of the simulation environment, the agents implemented, the progression of the simulation and the algorithm for anger and gratitude used by all emotional agents implemented (see sections 6.4.1, 6.4.2, 6.4.3 and 6.4.4 respectively).

6.4.1 Simulation Set-up

This section provides specific details regarding the simulation environment used to investigate the research questions listed in section 6.1. The fundamental features of this simulation are discussed in section 5.2.1 of this thesis so I will refrain from presenting this information again. Instead all information in this section is specific to the simulation version used in these experiments. In the simulations described in this chapter, two agents inhabit a virtual environment. The environment does not "wrap" because each agent can only interact with its opponent and each agent is capable of cooperating or defecting. There are no player/comparator sets implemented in this version of the simulation (as described in section 5.2.1 of chapter 5). Since I am looking to investigate if any of the emotional characters implemented may offer an improvement over any of Axelrod's strategies, it is sufficient for each agent to only play against one opponent rather than several.

6.4.2 Agent Details

To answer research question 4 from section 6.1: "*Do any of the emotional characters modelled offer an improvement upon the success exhibited by the TFT strategy identified as the most successful strategy in Axelrod's tournament? If so, why?*", some of the most successful strategies from Axelrod's tournament need to be implemented as agents so they are capable of playing in the simulation environment. Six types of agents are therefore modelled in this set of simulations (to be used in addition to the agents endowed with the emotional characters discussed in section 6.3) which do not require strategies, since their behaviour is entirely determined by their characters and their emotional states. Details of these Axelrod strategies along with their names are listed in table 6.2.

The behaviour of agents endowed with emotions is entirely determined by the agent's emotional character and current emotional state (see section 6.3) *only*; they do not use any of the Axelrod strategies above to determine their behaviour. Until an emotion is activated in an emotional agent, its behaviour is determined by its initial disposition,

TABLE 6.2: Descriptions of strategies from Axelrod's tournament that have been implemented.

Strategy	**Behaviour Description**
Mendacious	Always lies.
Veracious	Always tells the truth.
Random	Has a 1 in 2 chance of defecting/cooperating in each round.
TFT	Cooperates on the first round. In next round it mimics the opponent's behaviour from the previous round.
Tester	Defects on first round, if the opponent cooperates in the first round then the agent cooperates in rounds two and three but defects in round four (this cooperate, cooperate, defect cycle is then repeated for the remainder of the game). If, the opponent defects on the first round, the agent plays TFT for the rest of the game.
Joss	Plays TFT, but has a 1 in 10 chance of defecting on a round.

represented by a variable that serves to inform the agent as to whether it should cooperate or defect. Once an emotion has been elicited however, the variable no longer exerts an effect upon the agent's intentional behaviour. This imposes a further pseudo emotional character classification since it is now possible for an A1:G1 agent that initially cooperates to exist along with an A1:G1 agent that initially defects, for example. These initial behaviours need to be taken into account in the results analysis presented in section 6.5 since they may have considerable effects upon individual and total system scores.

6.4.3 Simulation Progression

A typical round in the simulations used in this chapter progresses as follows:

1. Non-emotional agents consult their strategy and emotional agents consult either their current emotional state if an emotion is active or their initial behaviour setting if not, and cooperate or defect accordingly with their opponent.

2. Agents calculate pay-offs.

3. If the agent is endowed with an emotional character its current emotional status is updated.

4. Round number is incremented by one and the environment is reset.

Pay-offs are used by Axelrod's notable strategies to inform their subsequent behaviour whereas emotional agents respond directly to the behaviour of the opponent in the past round. This is important as the point made in sections 6.2.1 and 6.2.2 is that anger and gratitude are fundamentally elicited in public goods games by the *behaviour* of an opponent rather than the perceived *distribution of pay-offs* (distribution of pay-offs may serve to inform emotion potential calculations). Each agent has an associated variable that keeps track of their individual score and a separate system variable keeps

track of the total system score (the sum of both agent's individual scores). Upon completion of the final round these values are recorded in an external text file. The total and individual scores received by agents are used to compare their success in exactly the same way as in Axelrod's tournament, this facilitates the acquisition of an answer to research questions 4 and 5 from section 6.1.

To provide data that can be used to answer research questions 4 and 5 from section 6.1, every agent will play against all others for 5 games of 200 rounds (as in Axelrod's tournament). It should be made clear that whenever a game ends after 200 rounds, any variables associated with the agents playing are reset so that the behaviour of agents in the preceding game does not affect the behaviour of agents in the new game.

6.4.4 Anger and Gratitude Algorithm

1. Agent x analyses its opponent y's behaviour in this round.
   (a) If y defected:
      i. If x's anger is currently active do not alter its anger potential towards y (x's anger remains active towards y).
      ii. Otherwise, x increases its anger potential towards y by 1.
   (b) If y cooperated:
      i. If x's gratitude is currently active do not alter its gratitude potential towards y (x's gratitude remains active towards y).
      ii. Otherwise, x increases its gratitude potential towards y by 1.
2. x checks anger and gratitude potentials towards y.
   (a) If x's anger potential towards y equals x's anger activation threshold, x's behaviour towards y is set to defect and x's anger and gratitude potentials towards y are reset to 0.
   (b) If x's gratitude potential towards y equals x's gratitude activation threshold, x's behaviour towards y is set to cooperate and x's anger and gratitude potentials towards y are reset to 0.
   (c) If x's anger or gratitude potentials towards y don't equal x's anger/gratitude activation threshold, x's behaviour is not altered.
      i. If x's anger or gratitude towards y is not active, x's behaviour towards y in the next round is determined by x's initial behaviour setting.
      ii. If x's anger is active, x will defect against y in the next round.
      iii. If x's gratitude is active, x will cooperate with y in the next round.

6.5 Results and Analysis

In this section I now address research questions 4 and 5 from section 6.1. In this investigation the definition of "success" is taken from Axelrod's definition of the term in the context of his tournament [7]. Ergo, in context of the simulations run in my own work, the most successful behaviour implemented by an agent is defined as the one that achieves the *highest total system score collectively against all agents.*

To measure the success of an agent the total system score or, more specifically, the *aggregated average total system score* (the sum of each average total system score achieved by an agent) is used. To maximise total system score two goals must be achieved:

- Cooperation must be *established quickly* between the members of the system.
- Cooperation must be *maintained* between the members of the system.

Ergo, a readiness to cooperate (responsiveness) and aversion to defection (tolerance) are both important factors. This is supported by a consideration of the results in table 6.3 where the initially cooperative A3:G1 agent - the most tolerant and most responsive - offers the greatest aggregated total average system score: 5895.6 of the emotional characters implemented. The TFT agent scores an aggregated average total system score of 5230.8 in comparison, so a more successful emotional strategy than the TFT strategy does indeed exist. Note however that this may not mean that A3:G1 is more successful than the TFT agent with respect to *every* agent it is pitted against.

This result is consistent with the observations of Jiang, Vidal and Huhns [95] and also Bazzan and Bordini [12] (see section 5.1 of chapter 5). With respect to Jiang, Vidal and Huhns, they state in [95] that, when compared with non-emotional agents, emotional agents always perform better since they can adapt their behaviour to a dynamic environment. Bazzan and Bordini in [12] on the other hand note that emotional characters increase cooperation rates among a population by up to 47% in their simulations. In this section however, I will dissect the performance of A3:G1 to determine why it is that this agent actually performs better than non-emotional agents and its other emotional counterparts, something that both Jiang, Vidal and Huhns and Bazzan and Bordini fail to do in their work.

To explain why A3:G1 is the most successful emotional character/strategy considered, its responsiveness and tolerance must be discussed in turn. Therefore, these criteria are discussed respectively in sections 6.5.1 and 6.5.2 (answering research question 4). I will then use the insights gleaned from identifying these criteria to present a further discussion of particularly interesting results obtained by the simulations in section 6.5.3 (answering research question 5).

TABLE 6.3: Aggregated average total system scores of all initially defecting/cooperative emotional characters.

	Initial Behaviour	
Character	*Coop*	*Defect*
A1:G1	5230.8	4344.8
A1:G2	5069.8	3634.6
A1:G3	4979.8	3599
A2:G1	5774.8	5061.8
A2:G2	5241.8	3707.2
A2:G3	5130	3659
A3:G1	5895.6	5167.4
A3:G2	5328.8	3744
A3:G3	5235.8	3702

6.5.1 Responsiveness and Total System Scores

Along with emotional characters A1:G1 and A2:G1, emotional character A3:G1 is one of the most responsive emotional characters implemented. Responsiveness or, the readiness of an agent to cooperate with an opponent, bears heavily upon the system's goal of establishing cooperation quickly in the iterated Prisoner's Dilemma game. This is because the quicker an agent is to reciprocate cooperation with cooperation, the more likely it is that concurrent cooperation will be established (important as the number of rounds in a game is finite).

The importance of responsiveness with respect to total system score is clearly observable in table 6.3. If emotional characters A3:G1, A3:G2 and A3:G3 are considered then it can be seen how, as responsiveness decreases, aggregated average total system scores decreases. The same pattern also holds true for emotional characters A1:G1-A1:G3 and A2:G1-A2:G3. The benefits of increased responsiveness are more pronounced if agents that periodically defect (random, tester and joss) are taken into consideration. The values of interest are displayed in table 6.4 where emotional character A3:G1 obtains the highest average total system score against each opponent. The same pattern also holds true for emotional characters A1:G1-A1:G3 and A2:G1-A2:G3 who initially defect or cooperate.

TABLE 6.4: Comparison of the average total scores (and standard deviations) for initially cooperative emotional agents with characters A3:G1, A3:G2 and A3:G3 when playing against Axelrod strategies that periodically defect.

	Opponent		
Emo. Ch.	*Random*	*Tester*	*Joss*
A3:G1	1002.8 (21.44)	1111 (0.00)	972.8 (177.45)
A3:G2	942 (27.21)	1089 (0.00)	488.8 (62.38)
A3:G3	902 (35.55)	1036 (0.00)	488.8 (62.38)

6.5.2 Tolerance and Total System Scores

It is not enough to simply establish cycles of cooperation though; to maximise the score of the system cooperation cycles must be maintained even when the other player temporarily defects (as self-interested agents will tend to do). In [7], Axelrod states that one of the fundamental properties of a successful strategy is that it is quick to punish. The problem with such behaviour is that it breaks cooperation cycles quickly if periodic defection occurs. This results in lower total system scores being achieved since constant DD may be locked into (if both players use the TFT strategy, for example). By one agent continuing to cooperate in the face of defection the system scores 5 rather than 2. If the defector then decides to cooperate again and is met with cooperation, a total system score of 6 is achieved. Therefore, by being more tolerant an agent must be prepared to suffer a reduction in its individual score.

As well as being one of the most responsive agents implemented in these simulations, A3:G1 is also one of the most tolerant along with A3:G2 and A3:G3. Therefore, the second part of its success is due to its ability to maintain cooperation cycles in the face of defection. To ascertain the effect of tolerance upon total system scores the aggregated average total system scores for A1:G1, A2:G1 and A3:G1 are compared (equally responsive but tolerance varies), these values can be found in table 6.3. As can be seen, as tolerance increases, the average aggregated average total system score increases. This trend holds irrespective of the emotional character's degree of responsiveness too i.e. if aggregated average total system scores for A1:G2, A2:G2 and A3:G2 or A1:G3, A2:G3 and A3:G3 are compared it can be seen how aggregated average total system score increases as the emotional character's degree of tolerance increases.

Like responsiveness, the benefits of increased tolerance are exemplified if the average total system scores of initially cooperative emotional characters A1:G1, A2:G1 and A3:G1 are considered when playing against the Axelrod strategies that periodically defect (random, tester and joss agents). Table 6.5 shows how an increased tolerance to defection increases average total system scores.

TABLE 6.5: Average total system scores (and standard deviations) of initially cooperative emotional characters A1:G1, A2:G1 and A3:G1 when playing against Axelrod strategies that periodically defect.

	Opponent		
Emo. Ch.	*Random*	*Tester*	*Joss*
A1:G1	900 (32.90)	1066 (0.00)	461.8 (38.15)
A2:G1	973.4 (20.33)	1100 (0.00)	895.4 (288.90)
A3:G1	1002.8 (21.44)	1111 (0.00)	972.8 (177.45)

6.5.3 Tolerance, Responsiveness and Individual Scores

In sections 6.5.1 and 6.5.2 above, I investigated the effects of responsiveness and tolerance upon aggregated average total system scores and average total system scores when

Axelrod strategies that periodically defect are played against thus providing an answer to research question 4 from section 6.1. Whilst these results are important, the effect of tolerance and responsiveness upon individual scores should also be discussed since success at the macro level does not entail success at the micro level. This section will therefore provide answers to research question 5 from section 6.1.

I begin with a consideration of how responsiveness affects the average individual scores of emotional agents when playing against the Axelrod strategies that periodically defect. Table 6.6 presents the relevant results and offers a number of interesting observations. Firstly, it can be seen how as responsiveness decreases, average individual scores for the emotional character agents considered, increases. The only exception to this trend is observed when emotional characters play against the joss agent. In this case, increasing responsiveness past moderate responsiveness does not have an effect. The trends related to average individual scores and responsiveness noted for emotional characters A3:G1, A3:G2 and A3:G3 also apply to emotional characters A1:G1-A1:G3 and A2:G1-A2:G3.

TABLE 6.6: Comparison of the average individual scores (and standard deviations) for initially cooperative emotional agents with characters A3:G1, A3:G2 and A3:G3 when playing against Axelrod strategies that periodically defect.

	Opponent		
Emo. Ch.	*Random*	*Tester*	*Joss*
A3:G1	372.4 (20.33)	443 (0.00)	449.4 (74.00)
A3:G2	417 (16.57)	487 (0.00)	239.4 (31.19)
A3:G3	446 (17.78)	513 (0.00)	239.4 (31.19)

Decreased responsiveness causes average individual scores to increase since emotional characters that are less responsive are capable of punishing opponents for longer periods following activation of anger. In this way, periodic defectors who hope to distribute the sucker's pay-off before re-establishing cooperation cycles (thus maximising both their individual score as well as the total system score) are punished more severely and the punisher recuperates its losses with interest. Furthermore, the quicker an agent is to re-establish cooperation, the more it can be taken advantage of. This form of "naivety" is disadvantageous from an individual standpoint but not from the system's perspective.

The average individual scores obtained by emotional characters A3:G1, A3:G2 and A3:G3 when playing against the joss agent are caused by a combination of the joss agent's TFT behaviour and the decreased responsiveness of emotional characters A3:G2 and A3:G3. Since emotional character A3:G1 is quick to cooperate it quickly re-establishes CC with a joss agent following periodic defection whereas emotional characters A3:G2 and A3:G3 will establish defection cycles. This increases the average individual score of emotional character A3:G1 because CC earns the individual more per round than DD. To clarify why the scores for emotional characters A3:G2 and A3:G3 plateau, consider the play histories in table 6.7. In round n the joss agent periodically defects and activates anger in all the emotional characters considered. This then causes the emotional

characters to defect in round $n+1$ but, because all emotional characters cooperated in round n, the joss agent cooperates in round $n+1$ (since it uses TFT). This is where the emotional characters diverge: in round $n+2$, emotional character A3:G1 will reciprocate the joss agent's cooperation in round $n+1$ due to its gratitude being activated whereas emotional character A3:G2 requires two cooperations (emotional character A3:G3 requires three). This establishes DD between the joss agent and emotional characters A3:G2 and A3:G3 since the joss agent reverts to TFT without activating gratitude in either of these agents and periodic cooperation is not possible. However, when playing against emotional character A3:G1, the joss agent will defect in round $n+2$ since it reciprocates A3:G1's defection from round $n+1$. Yet, in round $n+3$, both A3:G1 and the joss agent will re-establish CC since A3:G1 requires 3 defections to defect so the joss agent's defection in round $n+2$ has no immediate effect.

TABLE 6.7: How decreased responsiveness causes a reduction and plateau of average individual score when playing against agents that use TFT and periodically defect.

Round #	**A3:G1**	**Joss**	**A3:G2**	**Joss**	**A3:G3**	**Joss**
n	C	D	C	D	C	D
n + 1	D	C	D	C	D	C
n + 2	C	D	D	D	D	D
n + 3	C	C	D	D	D	D
Score	8	13	7	7	7	7

As stated, the trends observed for A3:G1, A3:G2 and A3:G3 also hold for A1:G1-A1:G3 and A2:G1-A2:G3 but the scores achieved are much lower. This is caused by the reduced tolerance of these emotional characters which, if the joss agent is primarily considered, causes DD to be locked into more quickly and lower individual scores to be obtained. It is the effect of tolerance upon individual scores that I will now turn my attention towards thus providing an in-depth discussion of how tolerance affects individual scores and why.

If the average individual scores for initially cooperative A1:G1, A2:G1 and A3:G1 are compared (see table 6.8) it can be seen that, in much the same way as responsiveness, as tolerance increases, emotional characters sacrifice individual score to maximise total system score. Also, as with responsiveness, the trend is reversed when these emotional characters play against the joss agent. However, the difference between tolerance and responsiveness is that increasing tolerance does not cause the individual score to plateau.

TABLE 6.8: Average individual scores(and standard deviations) of initially cooperative emotional characters A1:G1, A2:G1 and A3:G1 when playing against Axelrod strategies that periodically defect.

	Opponent		
Emo. Ch.	*Random*	*Tester*	*Joss*
A1:G1	449 (16.58)	533 (0.00)	228.4 (19.07)
A2:G1	398 (17.84)	465 (0.00)	417.2 (124.63)
A3:G1	372.4 (20.33)	443 (0.00)	449.4 (74.00)

The reversal of the trend outlined when A1:G1, A2:G1 and A3:G1 play against the joss agent (see table 6.8), occurs because of the joss agent's low likelihood of defection i.e. there is only a 10% chance that the agent will defect in any round. As tolerance increases, it is more likely that cooperation cycles will be maintained since the likelihood of the joss agent defecting twice consecutively is 0.01 (obviously, the likelihood of the joss agent defecting three times consecutively is even smaller; 0.001). So, the high tolerance of A3:G1 enables the highest likelihood of CC maintenance in the face of rare, one-off, periodic defections. In the case of A1:G1 however, periodic defection by the joss agent leads to an unending cycle of TFT play earning the emotional agent less over 2 rounds than it would if CC were re-established (5 instead of 6). This drawback of the TFT strategy is mentioned in section 2.4.3 of chapter 2.

I now consider scores that pertain to emotional characters A1:G1, A2:G1 and A3:G1 when pitted against the random agent to explain why the average individual score of each emotional character decreases as tolerance to defection increases. If an agent has a high probability of defecting (as the random agent does) then it is not unlikely (as it is for the joss agent) for the agent to defect a number of times consecutively. In such a situation, if an emotional agent is currently cooperating, the more tolerant the emotional agent is, the more often it will receive the sucker's pay-off. Thus, even though cooperation is maintained, the individual score of the tolerant agent suffers.

So far it has been experimentally verified that, as tolerance and responsiveness increase, the total system score increases whilst an agent's individual score decreases but how much of a trade-off between individual score and total system score occurs? To answer this I will consider tolerance; as tolerance increases, the difference between individual scores decreases, to clarify: emotional characters A1:G1 and A2:G1 have an average individual score difference of 50.8 whereas the difference between these values for emotional characters A2:G1 and A3:G1 is almost halved to 25.8. Using the Prisoner's Dilemma pay-off matrix (see table 2.1 in section 2.1.2 of chapter 2) the trade-off can be precisely calculated. For every point earned by the system in context of CC and DC or CD, an agent (in this case the emotional agent), must sacrifice 2 points with respect to its preferred individual score. This answer begs a further question: how much of a reduction in individual score is acceptable to achieve these system gains?

To determine when the trade-off between individual and system scores becomes unacceptable a number of *threshold* values must be calculated. Various maximal and minimal scores that can be achieved for or tolerated by each entity in the simulations needs to be considered to determine these thresholds. Table 6.9 identifies these values along with their method of calculation and maximum/minimum values:

The best possible score that any individual agent can achieve is 1000 whilst the worst is 0; achieved when a mendacious strategy plays against a veracious strategy (mendacious earns 1000, veracious earns 0). An individual score of 0 is the worst scenario possible for an agent yet, the lowest *acceptable* score from the point of view of a self-interested agent who always seeks to earn *some* pay-off is 200. A total system score of this amount can

TABLE 6.9: Threshold values present in the simulation with their method of calculation and maximum/minimum values.

Threshold Value	**Calculation**	**Max.**	**Min.**
Average Agent 1 Score (A1)	A1 Individual Score	1000	0
Average Agent 2 Score (A2)	A2 Individual Score	1000	0
Average System Score	A1 + A2	1200	400
Average Fairness Score	IF A1>A2 = A2/A1 ELSE A1/A2	1	0

only be achieved by two agents locking into DD for an entire game. The best possible total system score is 1200, achieved when two agents initially cooperate and maintain CC for a whole game. The worst total system score is achieved by two agents locking into a DD for a whole game, leading to a total system score of 400. Using a combination of the total system score and the individual scores of the agents in the simulation, a measure of fairness can be identified which ranges from 0 to 1. The closer the value is to 1 the more equal the two players' scores are. It is this measure that I will now focus upon.

The results obtained have so far established that an initially cooperative A3:G1 agent is more successful than an initially cooperative A1:G1 agent (the emotional version of the TFT agent). The reason for this success is due to A3:G1's increased *tolerance* i.e. its ability to maintain cooperation since the responsiveness of emotional characters A1:G1 and A3:G1 are equal. However, the total system scores produced by A3:G1 agents are not fairly distributed when compared to the fairness measures for equally responsive and tolerant characters when playing against Axelrod strategies that periodically defect (see table 6.10). For example: when playing against a random agent the average fairness value obtained by an initially cooperative A3:G1 agent is 0.59, whereas for an initially cooperative A1:G1 agent fairness is equal to 1 and for an initially cooperative A3:G3 agent fairness is equal to 0.98. Despite this, the system total achieved by an initially cooperative A3:G1 agent is much higher than that achieved by its less tolerant and less responsive peers (as previously discussed). It is conceivable that more tolerant agents that are highly responsive will produce greater total system scores at the expense of fairness, but only until a certain point i.e. when their individual score falls below the threshold of 200. Below this individual score the trade-off becomes unacceptable for a self-interested agent since consistent defection achieves a greater individual score.

Ratings of fairness for other emotional characters when grouped into responsiveness

classes are not considered but it is worth mentioning that emotional characters A1:G1, A2:G2 and A3:G3 all achieve the highest fairness ratings in the majority of cases when playing against Axelrod strategies that periodically defect. This is due to these emotional characters all being variations of the TFT strategy: A1:G1 cooperates/defects after receiving 1 cooperation/defection, A2:G2 cooperates/defects after receiving 2 cooperations/defections and A3:G3 cooperates/defects after receiving 3 cooperations/defections. If the fairness scores for these agents are considered when playing against all Axelrod strategies that periodically defect then it can be observed that A1:G1 always achieves the highest fairness rating out of the three emotional characters whilst A3:G3 always achieves the worst. This is because, as tolerance and responsiveness decreases, the emotional agents will punish and reward more slowly, creating a greater rift between the scores of the two agents playing.

TABLE 6.10: Average fairness ratings for emotional character agents with both types of initial dispositions when playing against random, tester and joss agents.

		Opponent		
Ini.Dis.	**Emo. Ch.**	*Random*	*Tester*	*Joss*
Coop	A1:G1	1	1	0.98
	A1:G2	0.67	0.5	1
	A1:G3	0.53	0.48	1
	A2:G1	0.69	0.73	0.87
	A2:G2	0.99	0.99	0.97
	A2:G3	0.80	0.81	0.97
	A3:G1	0.59	0.66	0.86
	A3:G2	0.79	0.81	0.96
	A3:G3	0.98	0.98	0.96
Defect	A1:G1	0.99	1	1
	A1:G2	0.64	1	0.98
	A1:G3	0.51	1	0.98
	A2:G1	0.7	1	0.88
	A2:G2	0.99	1	0.98
	A2:G3	0.78	1	0.98
	A3:G1	0.6	1	0.87
	A3:G2	0.81	1	0.98
	A3:G3	0.99	1	0.98

6.6 Chapter Summary

In this chapter I have attempted to model the emotions of anger and gratitude, justify the modelling decisions made, implement the basic emotional characters that will be used in the rest of this thesis and construct and run a number of simulations using agents endowed with these emotional characters in a bespoke environment. This bespoke environment pitted the emotional agents outlined against notable strategies from Axelrod's famous computer tournament [7] to allow me to provide answers to the

questions of whether any emotional characters are more successful than the TFT strategy (noted as the most successful strategy in Axelrod's tournament) and whether there are any interesting behavioural features associated with the base emotional characters proposed.

With respect to justification for the decision to model anger and gratitude, I believe that the case has been well made. Numerous pieces of literature were presented in section 6.2 that gave precise insights into the eliciting conditions, effects, probabilities of effect and intensity variables associated with anger and gratitude in the context of public goods games. From these pieces of literature I decided to implement these emotions so that anger would be elicited by, and would provoke defection in, an agent whilst gratitude would be elicited by and provoke cooperation. The effects of these emotions are guaranteed, as per the literature's descriptions. Furthermore, the potentials for these emotions are altered in the same way: when a cooperation or a defection is received, the potential for gratitude/anger is increased by one.

The question of when these emotions are elicited and exert some effect upon the intentional behaviour of the agent is answered in section 6.3. Anger and gratitude have three values that their activation threshold values may be set to: 1, 2 and 3 (as described in section 5.4.1 of chapter 5) and as such, nine emotional characters may be implemented. These emotional characters can be grouped according to their responsiveness and tolerance ratios and these ratios are the main focus of the investigation presented in the chapter.

In section 6.4 I gave an overview of the specifics of the simulation environment used to investigate the research questions listed in section 6.1 and outlined the order of play that the simulations follow. Section 6.5 then discussed some of the more interesting results obtained from running these simulations which provided an answer to the questions of whether or not any of the emotional characters implemented are more successful than the TFT agent when playing against other notable strategies from Axelrod's tournament and why. Furthermore, I have also examined whether there were any interesting behavioural features associated with the base emotional characters proposed. The key results to note are listed below:

- A1:G1 is the emotional analogue of the classical TFT agent due to its low tolerance and high responsiveness.
- The initially cooperative A3:G1 agent offers a significant improvement with respect to total system score than Axelrod's TFT agent for many reasonable initial configurations.
- The reason for A3:G1's success is its high *responsiveness* and *tolerance*. These qualities are especially important when playing against agents whose behaviour is uncertain such as the random, tester and joss agents.
- Increased tolerance ensures that cooperation cycles are maintained and system scores are still promoted by players.

- Increased responsiveness ensures that cooperation cycles are established quickly following defection from an opponent.
- A highly tolerant and responsive agent's individual score suffers since it may be taken advantage of more often and for longer periods of time.
- An agent's tolerance and responsiveness should never increase to the point where its individual score falls below 200, in this case DD is a better option as the agent will accrue at least *some* income.
- For highly responsive emotional character agents, increased tolerance decreases the fairness of income distribution since agents are able to be exploited for longer periods.

In the next chapter I turn attention to modelling the emotion of "admiration" and an analysis of its effects upon proliferating emotional characters throughout a larger population of entirely emotional agents.

Chapter 7

Admiration

In chapter 6, I experimentally verified that an initially cooperative A3:G1 agent is the most successful agent (with respect to total system score) out of the nine emotional characters and six Axelrod strategies implemented since it is both the most *tolerant* and the most *responsive*. Thus, this emotional character is slow to defect and quick to cooperate, qualities that facilitate both the *establishment* and *maintenance* of cooperation between itself and its opponent. In chapter 6, it was also shown that, whilst the increased tolerance of A3:G1 enables it to promote total system score most successfully, this benefit comes at a cost to individual score. The disparity between the individual scores of agents is termed as the *fairness* of the system. System fairness (as defined and calculated in this thesis, see section 6.5.3 and table 6.9 of chapter 6) ranges from 0 to 1, the closer to 0 the less fair the individual score distributions and vice-versa. Of its equally responsive counterparts (initially cooperative emotional characters A1:G1 and A2:G1), the initially cooperative A3:G1 agent achieved the lowest fairness scores when playing against agents that periodically defect (see table 6.10 in chapter 6).

So, given a population of emotional agents, would A3:G1 become the most prevalent emotional character if agents could directly adopt the emotional character of others? This question is the focus of this chapter and is primarily inspired by a consideration of Nowak and Roch's work on *spatial reciprocity* [136] discussed in section 2.4.1.3 of chapter 2. In [136], agents who play the iterated version of the Prisoner's Dilemma game are capable of altering their strategy if they observe that an opponent's individual score is greater than theirs. Under such conditions, Nowak and Roch found that this ability to switch strategies causes cooperation to readily proliferate throughout a population of individuals.

Given that the aim of the thesis is to investigate the effect of emotions rather than to find a "best" emotional character given certain initial conditions and having established that emotional agents can produce similar behaviour to Axelrods (in particular TFT, see chapter 6), it is not necessary to consider non-emotional agents further (one can always compare TFT with A1:G1, if necessary). In this chapter I therefore attempt to ascertain which emotional character (if any), becomes the most prolific in a population of purely emotional agents (one of which is analogous to TFT) if agents are capable of

comparing their scores against others and copying the emotional characters of those who are more successful than themselves?

Consequently, one of the main objectives of this chapter is to computationally implement an emotional class that bestows a mechanism which enables an agent, x, to copy the emotional character of an opponent, y if it is the case that y's individual score is greater than x's. The emotion class that is proposed to be modelled to enable such a mechanism is admiration: agents will be endowed with this emotion class in conjunction with the anger and gratitude emotion classes (since these emotions motivate the agent to intentionally cooperate or defect, see chapter 6). As a result, there are more emotional characters in this set of simulations than in the previous simulations detailed in chapter 6. Furthermore, the simulations themselves are radically altered since agents are all endowed with emotional characters (no strategies from Axelrods tournament are implemented) and simulations concern a population of agents rather than a pair. These modifications necessitate the introduction of the player and comparator sets discussed in section 5.2.1 of chapter 5 along with other factors.

The primary contribution of this chapter is to provide an investigation into, and show by way of experimental evidence, which emotional character(s) becomes prevalent in the population given different initial conditions. Also, is it the case that one or more emotional character(s) consistently become prevalent irrespective of initial conditions or is it the case that different emotional characters are selected for given different initial conditions? To find answers to these questions, I will again use a MAS test-bed. I also intend to make clear the eliciting conditions and effects of admiration so that I may computationally model the emotion class in a functional manner. This discussion should also benefit others who wish to implement the emotion class in a similar fashion. Thirdly, as a extension to the contribution made in chapter 6, the study undertaken in this chapter aims to show how anger, gratitude and admiration can be married together so that they may play a functional role in determining the intentional behaviour of agents in a simulated public goods game environment.

The structure of this chapter is largely similar to that of chapter 6: section 7.1 outlines the research questions that will be answered. Section 7.2 discusses admiration, specifically: justifications for its use and how it will be computationally modelled (eliciting conditions, effects, potential calculation and probability of effect). Section 7.3 presents the extended emotional characters implemented and used in the simulations whilst section 7.4 provides details of the simulations run. Section 7.5 analyses the simulation results and provides answers to some of the research questions outlined in section 7.1. The chapter is concluded in section 7.6 which presents a summary of the major conclusions drawn.

7.1 Research Questions

1. Is admiration a suitable emotion that can be used to enable agents to switch their emotional character if they are being outperformed by their opponents?

2. How should the emotion class of admiration be modelled computationally using the emotion model proposed in chapter 5, section 5.3?

3. How should the base emotional characters outlined in section 6.3 of chapter 6 be augmented with admiration?

4. What effect, if any, do various initial conditions have upon the prevalence of emotional characters in the simulations particularly:

   (a) What effect, if any, is exerted upon emotional character prevalence when the percentage of initial cooperators and defectors is altered?

   (b) What effect, if any, is exerted upon emotional character prevalence when admiration's activation threshold is increased or decreased?

   (c) What effect, if any, is exerted upon emotional character prevalence when different player and comparator sets are used?

5. Do any emotional characters become more prevalent overall, irrespective of initial conditions? If so, what are these emotional characters and why do they become prevalent?

Note that these research questions will produce results that have not (to the best of my knowledge), been explored by other computer-science researchers that have modelled emotions in public goods games. Therefore, it is impossible to comment upon whether any results are supported/refuted by other pieces of existing research. Essentially, the research performed in this chapter is entirely novel.

7.2 Why Admiration?

Admiration is defined by Ortony et al. in [142] as an emotion evoked by approving of someone else's praiseworthy action. Since I intend to implement a functional mechanism and emotion class inspired by emotion that will appraise the success of an opponent and mimic their emotional character if applicable, it would seem that admiration is a good candidate for such a role.

The decision to model and implement admiration in a MAS to spread successful characteristics was inspired by its occurrence in the OCC model [142], Steunebrink et al.'s logical formalism of the OCC model [186] and the observation that in both human and non-human societies, there are individuals who possess status or prestige and act as role-models. In this section I discuss research that considers the eliciting conditions

and effects of admiration in detail so as to produce a computationally functional implementation of this emotion class. This emotion class will be used by agents in a MAS simulation to provide answers to the research questions listed in section 7.1.

This section is split into two sub-sections: section 7.2.1 surveys existing literature that is concerned with the eliciting conditions and effects of admiration in human beings. This section also provides justifications for why admiration is used as a basis for the computational mechanism modelled that allows agents to switch their emotional character thereby addressing the first research question posed in section 7.1. Section 7.2.2 then provides the details of the emotion class of admiration i.e. its computational implementation, providing an answer to the second research question posed in 7.1.

As in chapter 6, the algorithm for admiration implemented in each emotional agent is not specified in detail until section 7.4.4 because additional information regarding the simulation set-up is required to fully understand its operation.

7.2.1 Eliciting Conditions and Effects

Like gratitude in section 6.2.2.1 of chapter 6, Algoe and Haidt's paper [5] would appear to provide one of the most rigorous discussions of admiration's eliciting conditions and effects. In the paper, admiration is posited to be elicited when individuals witness the extraordinary skills or talents of non-moral exemplars or when such individuals achieve some form of exemplary success. With respect to intentional behaviour, eliciting admiration in an individual provokes the admirer to emulate the actions of the admired, to "improve" themselves and to work harder to achieve their own goals. Therefore, Algoe and Haidt's definition of the emotion differs slightly from Ortony et al.'s definition since it is not just the action of another that elicits admiration; the *success* of another also has an input into whether or not the emotion is evoked. To clarify: instead of solely focusing upon any intrinsic merits of an action (such as how skilful the action is, for example), Algoe and Haidt consider admiration to focus on both the action and its consequences.

At this point it would be beneficial to note that in the rest of this discussion the term "inspiration" is used quite regularly, seemingly in place of "admiration"; this requires some explanation. In [5], Algoe and Haidt state that:

> "...admiration as we conceive it involves inspiration as its motivational output, driving the learning and relationship effects that we predict."

Inspiration would therefore seem to be a middle-man between admiration and its associated intentional behaviour. Consequently, since [5] is one of the most notable works regarding admiration, I use admiration and inspiration more or less interchangeably in this section because I am interested in the relationship between an emotion and the intentional behaviour it produces. If it is indeed the case that admiration excites inspiration which in turn provokes intentional behaviour then admiration and inspiration could be thought of as parts of a single process.

A wealth of experimental evidence is acquired and presented by Algoe and Haidt in their three studies (two of which have already been discussed in section 6.2.2.1 of chapter 6) to support the assertion that admiration's eliciting conditions are as described above; relevant details of these studies follow in the next section.

7.2.1.1 Algoe and Haidt's Studies

Study 1. With respect to the emotional word questionnaire used in this study (participants were asked to think of a specific time when they witnessed someone successfully overcoming an obstacle or handicap), the majority of words noted by participants were coded as either "enhancement", "acknowledgement" or "emulate". However, like their treatment of gratitude (discussed in section 6.2.2.1 of chapter 6), such evidence suffers from low psychological realism and potential fabrication by participants.

Study 2. Participants were asked to rate the degree to which a sports star had exceeded a "normal" person's talent, skill or ability on a 6 point scale from 0 (not at all) to 6 (very much) after watching a video of the sports star demonstrating their skill (study 2a). The mean response obtained with respect to this criteria was 5.74. Participants were also asked to rate how they they felt with respect to various emotion words on the same six point scale; the highest mean values were obtained for the words "admiration" and "respect". When these results are combined, strong evidence is provided to support the claim that admiration is indeed elicited when an individual witnesses an action that requires a particular degree of skill.

Furthermore, participants rated how motivated they felt with respect to ten items on a scale of -4 (much less) to +4 (much more) in study 2a and witnessing relevant events in everyday life in study 2b. For 2a, the two highest mean values obtained were for the items "*As a result of this event, I feel (more or less) like being like the other person.*" and "*As a result of this event, I feel (more or less) like achieving success.*". For study 2b, the two highest mean values obtained were again for the item "*As a result of this event, I feel (more or less) like being like the other person.*" but the second item shifts to "*As a result of this event, I feel (more or less) like telling others about the other person.*" With particular reference to study 2a, 28% of respondents descriptions stated that they wanted to emulate the sports star's actions, 43% stated that they wanted to engage in some physical activity and 35% reported that they wanted to engage in activities that would lead to some professional or academic success. Almost none of these descriptions were reported for the other emotions investigated in this study. These results would therefore indicate that whilst admiration causes individuals to abstractly aspire to emulate the actions of others, the emotion also practically motivates people to achieve success by emulating actions.

Study 3. Participants in the admiration control group were asked to write to someone they had regular interaction with about a time when that person displayed great skill or talent, for which they felt admiration. Following this, when participants were questioned, those in the admiration control group identified with the following

statements more so than participants in other emotion control groups: *"I would like to meet others who are like the person I wrote to"*, *"I would like to improve some aspect of myself"* and *"I have a desire to give something back to others"*. These results provide further evidence that admiration is elicited by witnessing actions that require great skill or talent and this emotion causes a desire to improve some aspect of oneself.

7.2.1.2 Other Studies of Admiration

Algoe and Haidt's work leads into considerations of other similar work that has been performed, all of which are relevant to the current discussion. Henrich and Gil-White in [87] argue that prestigious human individuals are the subject of admiration. In other words, prestigious individuals attract some degree of reverence due to some attribute that is perceived to be valuable to observers. As Henrich and Gil-White argue, what is valued could be any number of attributes: an individual may hold prestige due to possessing some discernible skill, some specific knowledge or simply because of their age. The authors suggest that admiration provokes *infocopying* of those who are admired. "Infocopying" is defined by Henrich and Gil-White as encompassing all methods that may be used to acquire information directly from others (observing the behaviour of others, questioning others etc.). Numerous pieces of experimental research are cited in support of this claim including Rosenbaum and Tucker's work [157] that illustrates the tendency of an individual to copy the choices of an highly competent individual with respect to betting on horses even when the individuals are betting on different races. The work of Kroll and Levy [106] is also cited; in their investigation the authors inadvertently discover that, when the performance and behaviour of top performing participants in multi-round, high-stake investment games are made public to other participants, their behaviour is emulated by others.

As mentioned in the discussion of Algoe and Haidt's work [5], it is not necessarily the action(s) of an agent that evoke admiration, rather an individual's evaluation of another's success with respect to some consequence of an action(s) may inspire such an emotion. This position is also argued for by authors such as Taylor and Lobel who observe that cancer patients prefer social interactions with patients whose prognosis and situation is better than their own since this provides them with greater inspiration [195]. Therefore, it may be that the outcome of their actions are admired but without the actions themselves there is no quality to admire. For example: if I earn a high individual pay-off this may be because I am slow to punish and quick to forgive and without these qualities the pay-off I achieve could be considerably different. In other words, there must be some notion of a causal link between the actions copied from someone that is admired and the success they achieve.

Lockwood and Kunda also investigate eliciting conditions and effects of admiration in [114] and [115]. The essential conclusion proposed by these works is that familiar prestigious individuals with valued skills elicit inspiration in those evaluating the individual but only if the valued skill is attainable by the evaluator. The amount of

literature fortifying such a claim is extensive and I would direct those interested towards [114]. Furthermore, Lockwood and Kunya also provide experimental evidence in [115] to assert that inspiration is undermined when an individual's own success resulting from an application of their own skills is highlighted in a similar situation. Therefore, when evaluating the success of others, evaluating the success of self is also important. With particular reference to [115], the authors present evidence to assert that the success of a superstar was aspired to since this success was based upon excellence in terms of tasks that participants were to tackle in the future.

Taking the research above into account it would seem that admiration is elicited by witnessing another perform an action(s) that require a discernible degree of skill to implement or entail some notable success due to their implementation. Furthermore, the action(s) performed and the success obtained must be attainable by the observer otherwise negative emotions are elicited. The effect of admiration in such circumstances is to motivate the admirer to become more like the admired individual and to emulate the behaviour of the admired individual. The empirical evidence that has been considered to enable me to make these assertions will be used as a basis for the computational modelling of the emotion class of admiration that is described in the next section.

7.2.2 Modelling Admiration Computationally

Given the discussion in section 7.2.1 the process of modelling admiration computationally as an emotion class is relatively straightforward. Algoe and Haidt's observations [5] are perhaps some of the most detailed and well argued insights into the eliciting conditions and effects of admiration and it is these observations that provide the foundations of the class. The first step is to decide what aspect of an agent another should admire. Since one of the concerns of this chapter is the investigation into whether or not one emotional character becomes prevalent overall or whether different initial conditions select for different emotional character prevalence, a way of determining an emotional character's *success* is required. As argued by Algoe and Haidt: when admiration is elicited in others, it is not because individuals solely admire the actions of another, but rather they admire the success that these actions impart. This claim is further strengthened by evidence provided by Taylor and Lobel [195], and Lockwood and Kunda [114], [115]. The success of an agent is therefore implemented as the eliciting condition of admiration in this work rather than any particular action of an agent.

In chapter 6, the success of an agent was determined by the total system score obtained by an agent playing with another in context of an iterated Prisoner's Dilemma game yet, given the discussions presented in section 7.2.1, admiration appears to be elicited by a consideration of an individual's success (as a consequence of their actions). Therefore, the eliciting condition of the computational model of admiration I propose is the *individual score* of an agent. However, in considering the individual scores of others, an agent must also consider its own score since Lockwood and Kunda in [115] argue that admiration of others is undermined when an individual is made aware of their own

success. Intensity variables that may affect admiration's potential are not discussed in section 7.2.1. Consequently, I will model the calculation and alteration of admiration's potential in the same way as I did for anger and gratitude (see section 6.2.2.2 of chapter 6). When an agent, x, observes that its opponent(s) have individual scores greater than x's, x will increase its admiration potential towards that opponent or those opponents by 1. Likewise, if x observes that it has achieved the highest individual score when compared to its opponents then it will increase its admiration potential towards itself by 1. The discussion of admiration in section 7.2.1 does not mention any situation where an individual may have felt admiration but masked its associated intentional behaviour. Admiration's probability of effect is therefore modelled in the same way as it is for anger and gratitude: when admiration is elicited in an agent, the agent is *guaranteed* to emulate the emotional character of the agent whom it admires, with no exceptions.

The consideration of admiration's effects in 7.2.1 indicates that when admiration is elicited in an individual, x, x desires to be more like the individual admired; y. Due to this, x will emulate aspects of y that it deems important to y's success. Since I am interested in investigating which emotional character gains prevalence in a population I have chosen to have agents emulate the admired agent's *base emotional character* i.e. the admirer will change its emotional character with respect to its anger and gratitude activation thresholds so they are equal to the admired agent's. It is important to note here that the current behaviour of the admired agent is not emulated i.e. if an agent, x, is cooperating and an opponent, y, is defecting, x will not begin to defect if admiration is elicited towards y. Thus an agent modifies its emotional character type, but not its current emotional state. This decision was taken since an agent's actions and current total score are ultimately determined by its emotional character rather than by its current behaviour. Therefore, the long-term determinant of success should be admired since the snapshot of the agent's current behaviour at the moment of admiration activation can not be said to determine its long-term success. A by-product of this decision is that agent pairs locked into DD/CC will not benefit or suffer from emotional character emulation whilst those locked in asynchronous cooperation-defection cycles will.

7.3 Emotional Characters

The discussion in this section addresses the third research question outlined in section 7.1, namely: "*How should the base emotional characters outlined in section 6.3 of chapter 6 be augmented with admiration?*". As in chapter 6, the intentional behaviour of emotional agents is driven by their emotional character. Specifically, the intentional behaviour of these agents is driven by a combination of:

1. The agent's current emotional potentials of anger and gratitude,
2. The agent's current emotional character which dictates the activation thresholds of the emotion classes implemented.

Therefore, simply endowing agents with admiration is not sufficient, this emotion must be combined with anger and gratitude so that agents are capable of producing intentional behaviour. Admiration will influence how this intentional behaviour is determined by altering the emotional character of the agent and thus the activation thresholds of the agent's anger and gratitude emotion classes. For the sake of consistency between emotion classes, admiration as a computational emotion is implemented in much the same way as anger and gratitude. Accordingly, there are three activation thresholds that the emotion may be set to: 1, 2 and 3 producing three extra emotional characters that each basic emotional character outlined in section 6.3 of chapter 6 may be augmented with: Ad:1, Ad:2 and Ad:3[1]. Since admiration is intended to augment the nine emotional characters that already exist, the total number of emotional characters possible to be implemented is now extended to a maximum of twenty-seven. This of course produces additional emotional characteristics; the adjective associated with admiration is *impressionable* (see section 5.4.1 of chapter 5). So, if the two-dimensional emotional character matrix from section 6.3 of chapter 6 (see table 6.1) is used as a basis, a third dimension may be added that extends each basic emotional character to include three extra emotional characters. The generic, basic emotional characters and their associated characteristics after being extended with admiration are outlined below for clarification:

- A*n*:G*n*:Ad:1 - Less/moderately/highly tolerant, less/moderately/highly responsive and highly impressionable.
- A*n*:G*n*:Ad:2 - Less/moderately/highly tolerant, less/moderately/highly responsive and moderately impressionable.
- A*n*:G*n*:Ad:3 - Less/moderately/highly tolerant, less/moderately/highly responsive and less impressionable.

Consequently, admiration is elicited in A*n*:G*n*:Ad:1 agents when they observe that either themselves or an opponent has obtained the highest individual score *once* out of all the agents considered. Admiration is elicited in A*n*:G*n*:Ad:2 agents when they observe that either themselves or an opponent has obtained the highest individual score *twice* out of all the agents considered. Finally, admiration is elicited in A*n*:G*n*:Ad:3 agents when they observe that either themselves or an opponent has obtained the highest individual score *three* times out of all the agents considered. The implications of altering these activation thresholds for admiration are discussed further in section 7.4.1 of this chapter.

7.4 Simulation Details

The simulation test-bed implemented in this chapter is fundamentally similar to that outlined in section 6.4 of chapter 6 i.e. an iterated Prisoner's Dilemma game is played

[1] **Ad** is a contraction for **Admiration**

Table 7.1: Number of agents with each emotional character in initial population.

Emo. Ch.	**# Agents With Emo. Ch. Initially**
A1:G1	38
A1:G2	37
A1:G3	38
A2:G1	37
A2:G2	38
A2:G3	37
A3:G1	38
A3:G2	37
A3:G3	38

in the context of a MAS simulation. The pay-off matrix illustrated in section 2.1.2 of chapter 2 (see table 2.1) is still used and the simulation environment is still based upon a two-dimensional grid. Other similarities still exist and will be discussed accordingly but there are some significant differences, some of which have already been touched upon in the introduction to this chapter, which will be discussed in section 7.4.1.

Particular details of agents i.e. initial behaviour determination, types of agents implemented and other actions performed will be discussed in section 7.4.2. Section 7.4.3 outlines the progression of the simulation both on a round-to-round and game-to-game basis. Finally, section 7.4.4 details the admiration algorithm implemented in all emotional agents used in the simulations.

7.4.1 Simulation Set-up

Although the environment of the simulation test-bed implemented in this chapter is still a two-dimensional space, the edges of this space now "wrap" unlike the simulation environment used in chapter 6. This feature has been implemented due to the requirement that all agents should be capable of comparing their individual scores to an equal number of opponents so as to not bias emotional character prevalence.

One of the most immediately noticeable differences in this version of the simulation compared to the version used in chapter 6 is that the number of agents implemented in the simulation has increased to 338. This number was chosen since it produced the largest possible space for the simulation view without having to resize the containing window. However, the decision to use 338 agents is a limitation of the simulation since it does not allow for an even number of emotional characters to be instantiated in the population. On reflection a number divisible by 9 would be better so that an equal number of emotional characters would be present in the initial population. This would ensure that there is as little influence upon emotional character prevalence as possible other than how successful an emotional character is given initial simulation conditions. The number of each emotional character used is given in table 7.1.

All agents in this simulation are now emotional i.e there are no agents who utilise Axelrod's strategies in this simulation (other than A1:G1 who produces behaviour that is analogous to the behaviour of the TFT strategy). The decision to increase the number of agents in the simulation and have an entirely emotional population was influenced by two observations noted in the discussion of Bazzan and Bordini's work [12] in section 4.3.7 of chapter 4. In their simulations, Bazzan and Bordini included 900 agents in their simulation to investigate how emotions may influence societal interactions. Given that I am looking to investigate whether an emotional character may be prevalent irrespective of particular initial conditions or not, it is necessary to include more than two emotional agents in the simulations implemented. If it is the case that only two emotional agents are located in the system, emotional character prevalence becomes diluted since the emotional character is only acted upon by one set of pressures (namely the intentional behaviour and emotional character of the other agent). By increasing the total number of agents, emotional characters are subjected to greater selection pressure since there are a wider array of emotional characters and intentional behaviours present in a local neighbourhood. This should result in emotional characters that impart long-term rather than periodic, short-term success, becoming more prevalent. Additionally, having a completely emotional population is required since all agents should be capable of comparing and switching to any emotional character if admiration is elicited in them. Note however, that this simulation environment differs to Bazzan and Bordini's since each agent is endowed with multiple emotion classes rather than just one. This ensures that societal interactions are richer and that agents behave in a more human-like and realistic fashion.

Related to the increase of total numbers of agents in the system is the inclusion of player and comparator sets, first discussed in section 5.2.1 of chapter 5. The intention of implementing these sets is to provide a means to control how much selection pressure emotional characters are placed under. Setting the player and comparator set to include more agents theoretically puts greater selection pressure upon an agent's emotional character. The reasoning behind this is as follows: by increasing the number of agents in the player (comparator) set it will be more likely that the agent will play (compare itself) against a greater variety of emotional characters. This should result in emotional characters that perform well against a larger subset of other emotional characters to become more prevalent within the population as a consequence. Note that the values for player and comparator sets are global i.e. if "parallel" is the value set for both the player and comparator set then every agent's player and comparator set is specified as such. There is no way of setting distinct player and comparator sets for individual agents which ensures that all agents play and compare themselves against an equal number of agents. The intention of this constraint is to prevent any bias from occurring with respect to emotional character prevalence. Player and comparator sets can however be set independently of each other i.e. in a given simulation the player set may be set to "parallel" for all agents whereas the comparator set may be set to "octagonal" for all

agents.

Selection pressure is also exerted upon emotional characters by alteration of activation thresholds for an agent's admiration emotion class. Setting the activation threshold of admiration higher means that emotional characters that promote individual success over a longer period of time will be more admired. If activation thresholds are set lower then emotional characters do not have to enable consistent individual success, periodic success is enough to promote emotional character prevalence throughout the population. Furthermore, since admiration is elicited by agents comparing their individual scores against others, it was deemed necessary to be able to control when agents would perform this comparison. If agents compare their scores after relatively few rounds then the plays enabled by specific emotional characters would not have time to develop making it easier or more difficult for emotional characters to be propagated throughout the population. Additionally, since the behaviour of agents is initially determined by their initial behaviour setting rather than their emotional state (see section 7.4.2), if agents compare their scores after too few rounds then the base emotional character may not have had time to exert any effect upon the agent's intentional behaviour and individual score. Therefore, judging the success of an emotional character based upon such information would be erroneous. For example: A1:G1's effects will be exerted on a round-to-round basis since it is quick to punish and quick to defect however, A3:G1's effects will be exerted after a greater number of rounds since whilst it is quick to cooperate, it is slow to defect. If the number of rounds between comparisons is increased then the effects of A3:G1 can be exerted fully whereas with fewer rounds, it is only the effects of A1:G1 that are relevant. It was therefore decided that agents should compare individual scores after five rounds have been played. This value was selected since it gives all emotional characters a chance to exert their full influence upon the intentional behaviour of agents so a fair and informed comparison can be made by other agents. The *comparison round* value is a constant and therefore, the effects of altering this number upon character prevalence was not explored.

The number of agents endowed with a particular basic emotional character, the number of less/moderately/highly impressionable agents, the number of initial cooperators/defectors, player/comparator set configurations may all be altered by specifying their values in an external file that is read by the simulation. This enables me to both provide answers for research questions 4a, 4b and 4c listed in section 7.1 and to keep initial conditions consistent across simulations.

7.4.2 Agent Details

Like the emotional agents implemented in chapter 6, the emotional agents in this simulation utilise an initial behaviour setting to generate their behaviour until either anger or gratitude is elicited. As also mentioned in section 7.4.1 above, it is possible to set the number of initial cooperators and defectors in much the same way as it is possible to set numbers of less/moderately/highly impressionable agents and numbers of basic

emotional characters. However, it is not possible to specify such values at an individual agent level, instead: once the desired numbers have been selected, each agent is randomly assigned an initial behaviour, admiration level and basic emotional character to prevent any bias upon upon character prevalence that may occur. For example: if it was possible to group together a set of initially cooperative, highly impressionable A3:G1 agents then this could seriously bias character prevalence.

7.4.3 Simulation Progression

So there is adequate opportunity for emotional character prevalence to emerge, each game in this version of the simulation consists of 500 rounds rather than 200 as in the previous simulation used in chapter 6. A typical game in the simulation proceeds as described below.

1. Agents are generated and randomly assigned an initial behaviour, admiration activation threshold and basic emotional character. The percentage of initial cooperators/defectors, less/moderately/highly impressionable agents and emotional characters in the population equals the percentages specified in the simulation's set-up files.

2. Agents consult their current emotional state if an emotion is active or their initial behaviour setting if not to determine whether or not to cooperate with or defect against each agent in its player set.

3. Each agent's emotional state is updated according to the opponent's behaviour in the last round and pay-offs are distributed by each agent as per pay-off distribution for the Prisoner's Dilemma (see table 2.1 located in section 2.1.2 of chapter 2).

4. If the number of rounds specified by the comparison round value have been played then agents compare the individual scores of all agents in their comparator set against each other and against their own individual score and update their admiration instances accordingly.

5. Round ends, return to step 3 and repeat until 500 rounds have been played.

At this point it is important to outline the mechanism used to determine what emotional character is copied if it is the case that two or more agents are equally admired (in other words, two instances of admiration are activated simultaneously). In this case, the admiring agent, x, will select the emotional character to switch to with equal probability i.e. if four emotional characters are admired then x will have a 1 in 4 chance of choosing each emotional character to switch to. However, this approach raises an issue caused by the randomness of initial emotional character assignment.

We wish to avoid multiple emotional characters of the same type in the comparator set influencing the emotional character copied, and so do not want to use a simple majority. Instead, to solve this issue, x assigns all admired emotional characters to a

"list" data structure. From here, all duplicate emotional characters will be removed so there is only one instance of an admired emotional character in the list and from here the probability-based method of emotional character selection is applied. This mechanism prevents an emotional character being selected due to more admired agents being endowed with it as a result of the emotional character distribution mechanism and ensures that emotional character prevalence is based upon the benefits imparted by the emotional character. Of course, alternative choices can be made, and the test bed could be used to explore them if desired.

To gather evidence to answer the research questions detailed in section 7.1, 14 different scenarios were run using the simulation constructed. Each of these scenarios sets initial percentages for:

- The percentage of initial defectors,
- The percentage of initial cooperators,
- The percentage of highly impressionable agents,
- The percentage of moderately impressionable agents,
- The percentage of less impressionable agents.

These scenarios and the percentages of each element listed above are presented in table 7.2 for reference. Note that in scenarios 1-5, the percentage of initial cooperators and defectors are altered whilst the percentage of the population that is highly, moderately and less impressionable is kept constant. Conversely, in scenarios 6-14, the percentage of initial cooperators and defectors is kept constant whilst the percentage of the population that is highly, moderately and less impressionable scenarios is altered. This design choice was made since research questions 4a and 4b necessitate it (see section 7.1).

Note also that since this chapter is concerned with admiration and how this emotion affects the prevalence of the nine emotional characters defined in chapter 6, modelling non-impressionable agents would not seem to add much to answering the question. If it is suspected that there might be significant results by including such agents, the test bed could be used to explore this additional question.

For each scenario there are 16 sub-scenarios run, each of which is repeated 10 times to harvest a data set large enough so that informed answers to research questions 4 and 5 (see section 7.1) can be offered. These sub-scenarios are concerned with the player/comparator set configurations: details of these sub-scenarios can be found in table 7.3. Player and comparator sets are altered in context of each scenario so that the effects of altering their configurations can be commented upon in a scenario-specific manner.

At the conclusion of every sub-scenario repeat, the simulation records the number of agents with each emotional character i.e. A1:G1 was represented by 45 agents at the end of the game, A1:G2 was represented by 50 agents at the end of the game etc.

TABLE 7.2: Scenario numbers and their associated percentages of initial cooperators/defectors and percentages of highly/moderately/less impressionable agents.

			Impressionability %		
Sc. #	**% Ini. Defectors**	**% Ini. Co-operators**	*High*	*Mod.*	*Low*
1	90	10	34	34	32
2	70	30	34	34	32
3	50	50	34	34	32
4	30	70	34	34	32
5	10	90	34	34	32
6	50	50	50	25	25
7	50	50	70	15	15
8	50	50	90	5	5
9	50	50	25	50	25
10	50	50	15	70	15
11	50	50	5	90	5
12	50	50	25	25	50
13	50	50	15	15	70
14	50	50	5	5	90

TABLE 7.3: Sub-scenario numbers and their associated player and comparator set configurations.

		Comparator Set			
		Parallel	*Orthogonal*	*Diagonal*	*Octagonal*
Player Set	*Parallel*	Sub-Sc.1	Sub-Sc.2	Sub-Sc.3	Sub-Sc.4
	Orthogonal	Sub-Sc.5	Sub-Sc.6	Sub-Sc.7	Sub-Sc.8
	Diagonal	Sub-Sc.9	Sub-Sc.10	Sub-Sc.11	Sub-Sc.12
	Octagonal	Sub-Sc.13	Sub-Sc.14	Sub-Sc.15	Sub-Sc.16

After all repeats of a sub-scenario are completed these values are then used in further calculations to provide answers for research questions 4 and 5 (see section 7.1). The ways in which these values are utilised are numerous and quite specific so discussions of this are postponed until section 7.5 where they are mentioned in their relevant contexts.

7.4.4 Admiration Algorithm

1. After 5 rounds, agent x puts its own individual score and the individual scores of all agent's in its comparator set into a list.

2. x ranks the scores from step 1 and determines the maximum.

3. x determines the emotional character(s) of agents who scored the maximum.

   (a) If the maximum score was obtained by one emotional character only, x increases the admiration potential associated with that emotional character by 1.

(b) If the maximum score was achieved by more than one distinct emotional character:

i. x assigns these emotional characters to a list.

ii. x removes duplicate emotional characters from the list and reorders list elements in sequential order.

iii. x selects a list element with a probability of $1/n$ where n = number of emotional characters after removing duplicates.

iv. x increases the admiration potential associated with the emotional character selected by 1.

4. x checks admiration potentials for all emotional characters.

(a) If any of x's admiration potentials for any emotional character equals x's admiration activation threshold, x changes its emotional character to that of the emotional character whose admiration potential equals x's admiration activation threshold. x then resets admiration potentials for all emotional characters to 0.

(b) If none of x's admiration potentials for all emotional characters does not equal x's admiration activation threshold, x does nothing.

7.5 Results and Analysis

The aim of this section is to provide answers to research questions 4 and 5 (see section 7.1). In section 7.4.3 it was noted that each of the 14 main scenarios simulated consists of 16 sub-scenarios with each sub-scenario being repeated 10 times. Consequently, a wealth of detailed information was acquired and due to the volume of results obtained, I have selected particular subsets of this data to present. I therefore provide a relatively brief summary of results in this section (compared to the volume of results acquired) that establish key points with regards to the research questions posed. In chapter 6, the nine base emotional characters were divided into three tolerance and responsiveness groups to aid explanation; the same practice will be adopted in this section. These tolerance and responsiveness groupings are reiterated below for reference:

- Tolerance groupings:
  - Low tolerance group: consists of emotional characters A1:G1, A1:G2, A1:G3.
  - Moderate tolerance group: consists of emotional characters A2:G1, A2:G2, A2:G3.
  - High tolerance group: consists of emotional characters A3:G1, A3:G2, A3:G3.
- Responsiveness groupings:

- Low responsiveness group: consists of emotional characters A1:G1, A2:G1, A3:G1.
- Moderate responsiveness group: consists of emotional characters A1:G2, A2:G2, A3:G2.
- High responsiveness group: consists of emotional characters A1:G3, A2:G3, A3:G3.

Before the results of the simulations are discussed in detail, it is necessary to first outline how the four possible play probabilities between an agent and its opponent CC,CD,DC and DD, can be altered by varying the scenario number (in context of scenarios 1-5). In this context the term "play" has a specific meaning and refers to the strategies employed by both agents in a single round. Considering such play probabilities is integral to understanding the conclusions drawn from an analysis of the results obtained since play probabilities have a great effect upon the interactions that develop between an agent and its opponents. Table 7.4 illustrates how initial play probabilities are affected by changing the simulation scenario. Scenarios 6-14 are not listed since the percentages of initial cooperators and defectors are kept constant at the values used in scenario 3. The fact that play probabilities have such a large effect upon interactions between agents in these simulations justifies the use of simulation as an experimental methodology. To analyse the effects of play probability in reality would be exceptionally complicated and time-consuming whereas, by using simulation, initial conditions and other variables that effect play probability can be controlled.

TABLE 7.4: The effect of scenario upon initial play probability.

Sc. #	**Prob. of DD**	**Prob. of DC**	**Prob. of CD**	**Prob. of CC**
1	0.81	0.09	0.09	0.01
2	0.49	0.21	0.21	0.09
3	0.25	0.25	0.25	0.25
4	0.09	0.21	0.21	0.49
5	0.01	0.09	0.09	0.81

A final important point to note with respect to these simulations relates to CC and DD plays. Once established between agents, these plays can not be destabilised since agents are not capable of periodic/random cooperation or defection. Therefore, the prevalence of emotional characters depends upon how emotional characters deal with CD/DC plays.

This section is broadly split into four sub-sections and since the results analysis is extremely detailed the headline results for each section are given in that sub-section's preamble. Sub-section 7.5.1 looks to provide an answer to research question 4a, sub-section 7.5.2 considers results that allow provides an answer to research question 4b and sub-section 7.5.3 answers research question 4c. The results considered in sub-section 7.5.4 aim to answer research question 5; because this section is relatively smaller and less detailed than previous sections, headline results will not be provided.

7.5.1 Emotional Character Prevalence and Initial Cooperation/Defection Percentages

Headline Results

- Tolerance
  - When initial defection is high in the population, a premium is placed upon increased tolerance since tolerance gives agents in initial CD/DC plays a greater chance of establishing CC plays with opponents (earning them a higher individual score than what would be obtained through establishing DD plays with opponents).
  - As the percentage of initial cooperators in the population increases, a premium is placed upon reduced tolerance since initial defectors have a greater chance of exploiting initial cooperators. Since the sucker's pay-off yields the least preferred outcome for the "sucker", intolerance is valued because it reduces exploitation by initial defectors.
- Responsiveness
  - When initial defection is high in the population, a premium is placed upon highly responsive emotional characters. These emotional characters have the highest likelihood of establishing CC plays from CD/DC plays thus reducing the likelihood that the sucker's pay-off is received or DD is established
  - As the percentage of initial cooperators in the population increases, a premium is placed upon moderately responsive emotional characters. This is due to their ability to exploit opponents for one more round before establishing CC when compared to highly responsive emotional characters. This additional round of exploitation earns the moderately responsive agent an increased individual score resulting in a greater likelihood of this emotional character being admired. This increased likelihood of admiration increases the potential prevalence of this emotional character in the population.

Methodology To investigate whether there is an effect upon emotional character prevalence due to differing percentages of initial cooperators and defectors (research question 4a, see section 7.1), the number of agents endowed with each emotional character following all 500 rounds of a game being played are recorded. This leads to the construction of a table like the example shown in table 7.5.

The information contained in tables like 7.5 are used to determine the *placing* of each emotional character with respect to that sub-scenario repeat. This calculation simply involves ordering emotional characters based upon the number of agents with that emotional character in the final population. The emotional character represented by the highest number of agents in the final population places first, for example. If

Table 7.5: Example of emotional character representation table derived from one repeat of a sub-scenario.

Emo. Ch.	**# Agents in Final Pop.**
A1:G1	38
A1:G2	29
A1:G3	46
A2:G1	30
A2:G2	30
A2:G3	35
A3:G1	45
A3:G2	51
A3:G3	34

two emotional characters are represented by an equal number of agents in the final population then those emotional characters will tie in the relevant position. If table 7.5 is used then a table like 7.6 is produced by this ordering.

Table 7.6: Example of emotional character placing derived from one repeat of a sub-scenario.

Placing	**Emo. Ch.**
1st	A3:G2
2nd	A1:G3
3rd	A3:G1
4th	A1:G1
5th	A2:G3
6th	A3:G3
7th	A2:G1, A2:G2
8th	A1:G2
9th	N/A

By using tables such as 7.6, I am able to calculate the frequency of placing for each emotional character over the entire sub-scenario by summing together the number of times each emotional character placed in each position for each of the sub-scenario's repeats. If table 7.6 is used alone as a basis for an example then the table produced would look something like table 7.7. The final table produced for each sub-scenario is essentially the same as table 7.7 but is based upon ten 7.6 tables.

Tables such as 7.7 are then used to calculate the total frequency with which an emotional character places in each position over the entire scenario. To do this, the total number of times an emotional character places in each position for each sub-scenario is summed together. By performing these calculations for scenarios 1-6 the effect of initial cooperation and defection rates upon emotional character prevalence can be established since the independent variables in these scenarios are the percentage of initial cooperators and defectors in the population. For the sake of readability and brevity I will restrict discussions of effects to a consideration of tolerance (see section

TABLE 7.7: Example of total emotional character placing for a sub-scenario.

	Placing Frequency								
Emo. Ch.	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
A1:G1	0	0	0	1	0	0	0	0	0
A1:G2	0	0	0	0	0	0	0	1	0
A1:G3	0	1	0	0	0	0	0	0	0
A2:G1	0	0	0	0	0	0	1	0	0
A2:G2	0	0	0	0	0	0	1	0	0
A2:G3	0	0	0	0	1	0	0	0	0
A3:G1	0	0	1	0	0	0	0	0	0
A3:G2	1	0	0	0	0	0	0	0	0
A3:G3	0	0	0	0	0	1	0	0	0

TABLE 7.8: Frequency of placing in positions 1-3 in context of scenarios 1-5 for tolerance groups of emotional characters.

Sc. #	**A:1**	**A:2**	**A:3.**
1	80	173	253
2	56	140	309
3	80	126	297
4	147	140	216
5	225	125	167

7.5.1.1) and responsiveness (see section 7.5.1.2) rather than analysing the performance of each emotional character for each scenario (such information would be superfluous to this discussion and would take a great deal of time to explain). Furthermore, by grouping emotional characters into their relevant tolerance and responsiveness characteristics it is possible to comment upon whether initial cooperation and defection percentages select for particular responsiveness or tolerance groupings.

In determining what effect the percentage of initial cooperators and defectors has upon the prevalence of particular tolerance and responsiveness emotional character groups, I have calculated the total frequency that these particular groups have placed in positions 1-3. Considering these placings in isolation facilitates analysis and discussion whilst also providing a fair representation of emotional character prevalence since emotional characters that have placed most frequently in positions 1-3 can be said to be significantly prevalent.

7.5.1.1 Effects Upon Tolerance

General Trends Varying the initial percentage of cooperators and defectors in the population would appear to have a great effect upon tolerance group prevalence as shown in table 7.8 and figure 7.1. A discussion of the general trends is found below with each scenario from 1 to 5 considered in turn.

Generally speaking it can be asserted that: as the percentage of initial cooperators increases in the population, high tolerance is admired less whilst lesser tolerance is

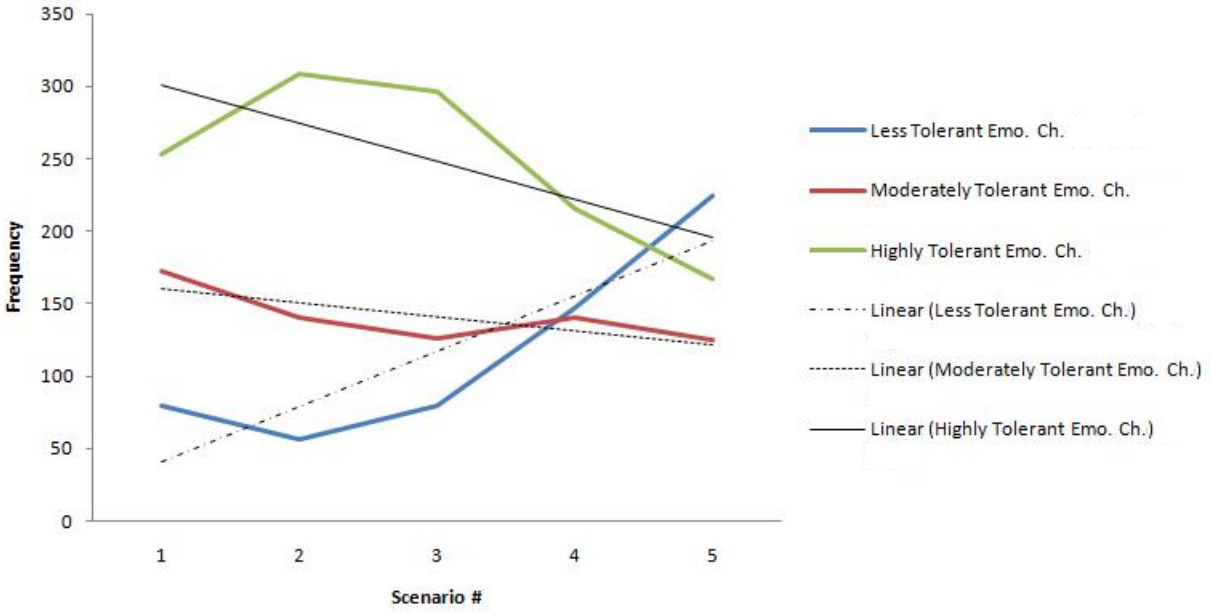

FIGURE 7.1: Frequency of placing in positions 1-3 in context of scenarios 1-5 for tolerance groups of emotional characters.

admired more (gradient of trend lines for highly and less tolerant emotional characters equal -26.5 and +38.1 respectively). Moderate tolerance is admired less as more initial cooperators are introduced however, this reduction in admiration is not as pronounced as that for highly tolerant emotional characters (trend-line gradient equals -9.6 compared to -26.5).

In considering tolerance I will focus upon agents that initially cooperate since tolerance only has an effect upon how quickly the agent responds to defection from another with defection. Therefore, considering agents that initially defect in context of tolerance is irrelevant because tolerance does not have any direct effect upon when the agent will cooperate.

Determination of Prevalence by Tolerance The prevalence of tolerant emotional character groups is determined by the factors listed below. These factors are listed in order of how much they promote emotional character prevalence together with a reason why the factor affects prevalence more/less than the other factors listed.

1. *Avoiding the establishment of DD plays from CD plays.* Once established, DD plays cannot be broken resulting in low individual scores being earned by both agents involved in the play for the remainder of the game.

2. *The ability to establish CC plays from CD plays.* Once established CC plays cannot be broken resulting in respectable individual scores being earned by both agents involved in the play for the remainder of the game.

TABLE 7.9: Play histories of an initially cooperative, highly tolerant agent when faced with less, moderately and highly responsive emotional agents that initially defect.

Round #	**A:3**	**G:3**	**A:3**	**G:2**	**A:3**	**G:1**
1	C	D	C	D	C	D
2	C	D	C	D	C	C
3	C	D	C	C	C	C

TABLE 7.10: Play histories of an initially cooperative, moderately tolerant agent when faced with less, moderately and highly responsive emotional agents that initially defect.

Round #	**A:2**	**G:3**	**A:2**	**G:2**	**A:2**	**G:1**
1	C	D	C	D	C	D
2	C	D	C	D	C	C
3	D	D	D	C	C	C

3. *The ability to avoid the maintenance of symmetric/asymmetric CD/DC plays from CD plays.* If a CD/DC play is met upon and CC is not established then consequential play histories can be highly variable. In such a situation, agents may establish a TFT play history or a far more complex CD/DC play history may emerge. Given this insight, it is preferable for agents to establish CC since the relevant agents are guaranteed to earn 3 points per round whereas the points earned by individual agents in a variable CD/DC play history are potentially less consistent.

With respect to the tolerance groups themselves, each group handles CD/DC plays in a characteristic way and as stated above, it is how these plays are handled that results in the prevalence values observed in table 7.8 and figure 7.1.

In the context of the emotional characters used in these simulations and the simulation set-up, highly tolerant emotional characters never immediately establish DD plays from CD/DC plays since there are no emotional characters capable of continuing to defect after having received three cooperations. In other words, there are no emotional characters whose activation threshold for gratitude is greater than A3:G1, A3:G2 or A3:G3's activation threshold for anger. Highly tolerant emotional characters also have a greater likelihood (2/3) of establishing CC plays from CD/DC plays rather than maintaining stabilised or destabilised CD/DC plays (1/3). Table 7.9 illustrates these features of highly tolerant emotional characters by taking into consideration play histories of highly tolerant agents that initially cooperate when playing against less/-moderately/highly responsive emotional characters that initially defect.

Moderately tolerant emotional characters have an equal probability (1/3) of maintaining stabilised or destabilised CD/DC plays; establishing CC from CD/DC plays; establishing DD from CD/DC plays. The play histories that can occur when an initially cooperative and moderately tolerant emotional character meets a less/moderately/highly responsive emotional character are shown in table 7.10.

TABLE 7.11: Play histories of an initially cooperative, less tolerant agent when faced with less, moderately and highly responsive emotional agents that initially defect.

Round #	**A:1**	**G:3**	**A:1**	**G:2**	**A:1**	**G:1**
1	C	D	C	D	C	D
2	D	D	D	D	D	C

When faced with an opponent that initially defects, an initially cooperative and less tolerant emotional character has a 2/3 chance of establishing DD or a 1/3 chance of maintaining a stabilised or destabilised DC/CD play. Indeed, one of Axelrod's rules of success namely, "punish defection immediately with defection" [7], appears to work counter-productively in this instance. Such a result is not novel (see Nowak and Sigmund's research into the GTFT strategy [137]) however, the fact that this rule is challenged by a population of agents using strategies inspired by human emotion is notable. Table 7.11 illustrates how these play histories emerge when an initially cooperative, less tolerant agent plays against a less/moderately/highly responsive emotional character that initially defects.

Scenario 1 Since the majority of the population lock into DD plays initially and the probability of initial CC plays is so unlikely (see table 7.4), there is a great premium placed upon emotional characters capable of performing well with respect to the three factors listed previously since the majority of agents obtain very low individual scores.

In the context of scenario 1, highly tolerant emotional characters are most prevalent since they can never establish DD plays from initial CD/DC plays and they have a greater likelihood of establishing CC plays rather than maintaining stabilised or destabilised CD/DC plays (see table 7.9).

Moderately tolerant emotional characters are less prevalent than highly tolerant emotional characters but more prevalent than less tolerant emotional characters since they have an equal chance of establishing CC/defection and maintaining stabilised/destabilised CD/DC plays with opponents (see table 7.10).

Less tolerant emotional characters are least prevalent in scenario 1 for the reasons discussed earlier: it is impossible for these emotional characters to establish CC from initial CD/DC plays, there is a 2/3 chance that DD will be established from CD/DC plays and a 1/3 chance that stabilised/destabilised CD/DC plays will be maintained. This results in less tolerant emotional characters achieving some of the lowest individual scores possible.

Scenario 2 Scenario 2 sees an increase in the initial percentage of cooperators and a decrease in the initial percentage of defectors within the population. With reference to table 7.4, the likelihood of agents meeting upon DD plays initially decreases from 0.81 to 0.49, the likelihood of agents meeting upon CC increases from 0.01 to 0.09 and the likelihood of CD plays occurring initially increases from 0.09 to 0.21. The salient point

to note is that the likelihood of DD being met upon initially is still higher than that of CC or CD/DC being met upon initially.

Since CC and CD/DC plays are more likely initially, there is a premium placed upon the ability to avoid the maintenance of stabilised/destabilised CD/DC plays and to establish CC plays from initial CD/DC plays (since agents in CD/DC plays must try to match the individual scores obtained by agents who meet upon CC initially to become prevalent). As previously mentioned, highly tolerant emotional characters have the greatest probability (2/3) of establishing CC with initial defectors out of all three tolerance groups identified and only a 1/3 chance of maintaining stabilised/destabilised CD/DC plays resulting in their continued prevalence.

Moderately and less tolerant emotional characters are less likely to establish CC plays when met with CD/DC plays (1/3 and 0/3 respectively rather than 2/3 for highly tolerant characters). This results in reduced individual scores when compared to highly tolerant agents and consequently such emotional characters are admired less and are therefore less prevalent. The likelihood of maintaining stabilised/destabilised CD/DC plays is equal for all tolerance groups (1/3) so the reduced prevalence of less tolerant agents when compared to moderately tolerant agents must be due to their increased likelihood of establishing DD cycles from CD/DC plays (2/3 compared to 1/3).

Scenario 3 The probability of encountering a CD/DC play initially is at its highest in this scenario (0.25, see table 7.4). CC plays are also more likely to be established initially (0.25) due to the 20% increase of initial cooperators in the population. Consequently, there is less of a premium placed upon the ability to establish CC with opponents since it is easier to come by initially thus, the reduced prevalence of highly tolerant emotional characters is observed. Since the ability of an emotional character to maintain cooperation is never admired, the ability to avoid excessive *exploitation* by defectors in CD/DC plays would appear to have a premium placed upon it in scenario 3.

The increased prevalence of less tolerant characters in scenario 3 is supported by the reasoning supplied above. Exploitation of less tolerant emotional characters is only ever possible for one round compared to two rounds for moderately tolerant agents and three rounds for highly tolerant agents. Therefore, intolerance towards receiving the sucker's pay-off becomes beneficial and is subsequently admired. This further explains the reduced prevalence of highly and moderately tolerant emotional characters.

Scenario 4 Scenario 4 is the first scenario considered where the probability of establishing CC initially is higher than the probability of establishing any other type of initial play (see table 7.4). Accordingly, increased tolerance is devalued further since there is less of a premium placed upon establishing CC from CD/DC plays because it is so likely to occur initially. Instead, intolerance has a greater premium placed upon it since initial defectors have more of a chance of exploiting initial cooperators in this scenario when

compared to scenario 3. Therefore, an increase in the prevalence of less tolerant emotional characters and a decrease in the prevalence of highly tolerant emotional characters is observed.

The increase in prevalence of moderately responsive emotional characters would appear to be explained by the increase in prevalence of emotional characters A2:G2 and A2:G3 which would indicate that the benefits conferred by reduced responsiveness are now being selected for as well. These effects will be discussed in more detail in section 7.5.1.2.

Scenario 5 Continuing from scenario 4, intolerance is a quality that is admired more in this scenario than ever before due to the exceptionally high likelihood that CC will be met upon initially. However, since defection is so sparse initially the prevalence of emotional characters is driven by the benefits imparted by responsiveness.

7.5.1.2 Effects Upon Responsiveness

General Trends As explained in the introduction to section 7.5.1.1, altering initial percentages of cooperators and defectors in the population has the greatest effect upon tolerance groupings but there still remains some interesting points to note with respect to the effects exerted upon responsiveness groups. Table 7.12 and figure 7.2 summarise the effects numerically and graphically and, as with section 7.5.1.1, a discussion of general trends will be provided before exploring the specific details of each scenario.

TABLE 7.12: Frequency of placing in positions 1-3 in context of scenarios 1-5 for responsiveness groups of emotional characters.

Scenario #	**G:1**	**G:2**	**G:3**
1	202	197	107
2	225	174	106
3	239	178	86
4	213	193	97
5	174	199	144

The point made regarding the reduced effect of initial defector and cooperator percentages upon responsive emotional character grouping prevalence compared to tolerant character grouping prevalence is illustrated by the flatter trend-lines displayed in figure 7.12 when compared to figure 7.8. Generally, admiration trends differ slightly from those that exist for tolerance groupings (see section 7.5.1.1) since moderately responsive emotional characters become more prevalent as the percentage of cooperators in the initial population increases (trend-line gradient equals +2.3).

The same general trends hold true with respect to highly responsive/tolerant and less responsive/tolerant emotional character groupings in relation to prevalence trends for tolerance groups i.e. highly responsive emotional characters become less prevalent and less responsive emotional characters become more prevalent as a greater percentage of the initial population cooperates. However, the respective decreases and increases in

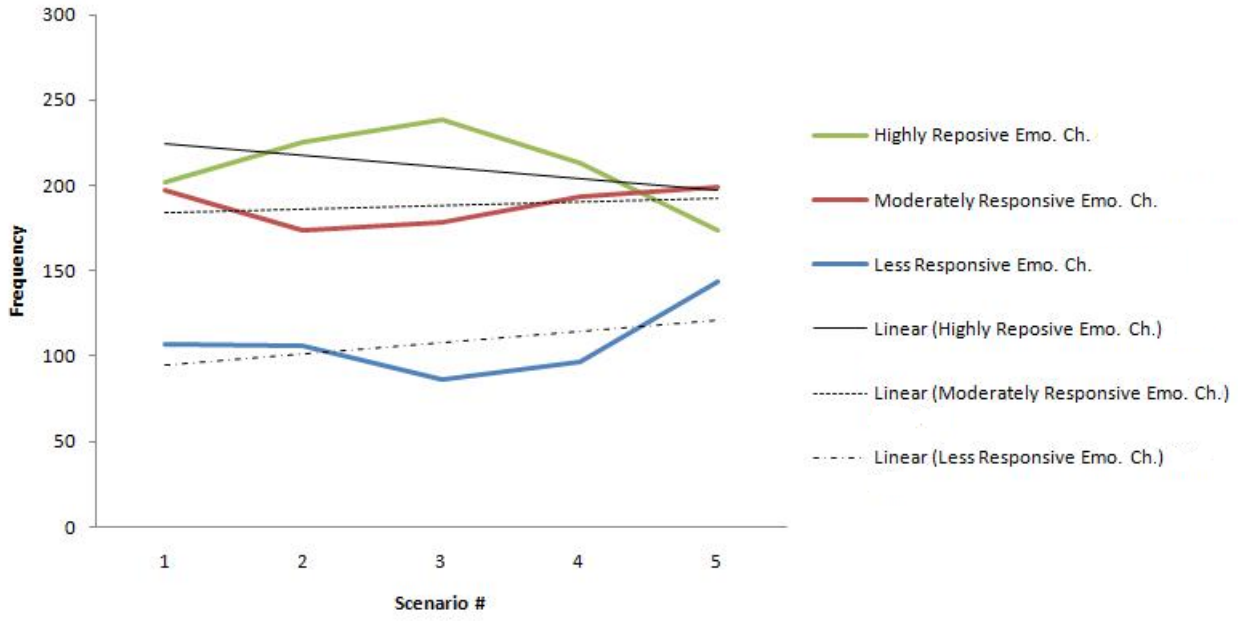

FIGURE 7.2: Frequency of placing in positions 1-3 in context of scenarios 1-5 for responsiveness groups of emotional characters.

prevalence are much less pronounced when compared with equivalent tolerance groups; trend-line gradients for highly and less responsive emotional characters equal -6.8 and +6.5 respectively. Compared with the trend-line gradients for highly tolerant emotional characters (-26.5) and less tolerant emotional characters (+38.1), the difference is clear.

In considering responsiveness I will focus upon agents that initially defect since responsiveness has an effect upon how quickly the agent responds to cooperation from another. Therefore, considering initially cooperative agents in context of responsiveness is irrelevant because responsiveness does not have any effect upon when the agent will defect when faced with defection.

Determination of Prevalence by Responsiveness In section 7.5.1.1 it was shown how emotional character prevalence is determined by the ability of each tolerance group to deal with the occurrence of CD/DC plays with respect to three factors (avoiding the establishment of DD, avoiding the maintenance of symmetric/asymmetric CD/DC, promoting the establishment of CC). In the same vein, different responsiveness groupings also have particular ways of dealing with these plays and are thus evaluated in accordance with these factors. Note however that the second factor is augmented with *exploitation*; responsiveness dictates how quickly an emotional character cooperates when defecting so, if an emotional character can exploit an opponent for longer before establishing CC, this will increase the individual score of agents endowed with this emotional character making it more likely that the emotional character will be admired.

1. *Avoiding the establishment of DD plays from DC plays.*

2. *The ability to exploit opponents before establishing CC plays from DC plays.*

TABLE 7.13: Play histories for highly responsive agents that initially defect when faced with initially cooperative less, moderately and highly tolerant emotional agents.

Round #	**G:1**	**A:1**	**G:1**	**A:2**	**G:1**	**A:3**
1	D	C	D	C	D	C
2	C	D	C	C	C	C

TABLE 7.14: Play histories for moderately responsive agents that initially defect when faced with initially cooperative less, moderately and highly tolerant emotional agents.

Round #	**G:2**	**A:1**	**G:2**	**A:2**	**G:2**	**A:3**
1	D	C	D	C	D	C
2	D	D	D	C	D	C
3	D	D	C	D	C	C

3. *The ability to avoid the maintenance of symmetric/asymmetric CD/DC plays from DC plays.*

Highly responsive emotional characters never immediately establish DD plays from DC/DC plays since their activation threshold for gratitude is less than or at least equal to any other emotional character's activation threshold for anger. Highly responsive emotional characters also have a greater likelihood (2/3) of establishing CC plays from CD/DC plays rather than maintaining stabilised or destabilised CD/DC plays (1/3). Table 7.13 illustrates these features of highly responsive emotional characters by taking into consideration play histories of highly responsive agents that initially defect when playing against less/moderately/highly tolerant emotional characters that initially cooperate.

Moderately responsive emotional characters have an equal probability (1/3) of maintaining stabilised or destabilised CD/DC plays; establishing CC plays from CD/DC plays; establishing DD plays from CD/DC plays. Moderately tolerant emotional characters are also capable of exploiting opponents for one more round than highly responsive emotional characters whilst still retaining the ability to establish CC afterwards (increasing individual score and guaranteeing the acquisition of 3 points per round for the remainder of the game). Whilst less tolerant emotional characters are capable of exploiting opponents for three rounds (one round more than moderately responsive emotional characters), this extra round of exploitation results in DD being established thus resulting in a reduced individual score. The key point here is that the exploitation of moderately responsive agents can be described as being in a *goldilocks*[2] zone. The play histories that can occur when an initially cooperative and moderately tolerant emotional character meets a less/moderately/highly responsive emotional character are shown in table 7.14 for clarification.

[2]Agents are neither too responsive resulting in establishing CC too early nor too unresponsive resulting in the establishment of DD. Their degree of responsiveness is *just right*.

TABLE 7.15: Play histories for less responsive agents that initially defect when faced with initially cooperative less, moderately and highly tolerant emotional agents.

Round #	**G:3**	**A:1**	**G:3**	**A:2**	**G:3**	**A:3**
1	D	C	D	C	D	C
2	D	D	D	C	D	C
3	D	D	D	D	D	C
4	D	D	D	D	C	D

Less responsive emotional characters that initially defect have a 2/3 chance of establishing DD or a 1/3 chance of maintaining a stabilised/destabilised DC/CD play from a CD/DC play. Table 7.15 illustrates how these play histories emerge.

Scenario 1 Since initial CC plays are so unlikely in this scenario (0.01, see table 7.4), there is a high premium placed upon the ability to establish CC from DC/CD plays. This results in the prevalence pattern observed in table 7.12 and figure 7.2 because highly responsive emotional characters have a 2/3 chance of establishing CC from CD/DC plays, moderately tolerant characters have a 1/3 chance of achieving this and less tolerant characters are not capable of establishing CC from CD/DC plays.

Scenario 2 In scenario 2 there is an observed increase in the prevalence of highly responsive characters whilst the prevalence of moderately responsive characters decreases (the prevalence of less responsive emotional characters remains the same for the same reasons described in context of scenario 1). Upon first consideration this may seem paradoxical given that, when faced with an initial cooperator, moderately responsive emotional characters that initially defect are capable of obtaining a higher individual score than highly responsive emotional characters that initially defect (due to their ability to establish CC following exploitation of an opponent). Furthermore, when compared with scenario 1, there is more of a likelihood that an opponent that cooperates initially is met. However, this is only true when a highly tolerant emotional character that initially cooperates is played against. Whilst moderately responsive emotional characters are in the so-called "goldilocks" zone with respect to exploitation of opponents, they run a significant risk of earning less than their highly responsive contemporaries if they play against moderately tolerant emotional characters. In this case, highly tolerant agents will exploit the moderately tolerant opponent then lock into CC whereas moderately responsive agents will exploit the moderately tolerant opponent and then establish a CD/DC play which can result in a number of different play histories. Depending upon the spatial distribution of opponent emotional characters throughout the population, it may well be that this gamble does not always pay off and therefore the safer exploitation of highly responsive emotional characters is admired instead.

Scenario 3 In scenario 3, the prevalence of less responsive emotional characters is noticeably reduced (see table 7.12 and figure 7.2). This observation is explained by considering why highly and moderately responsive emotional characters increase in prevalence. Essentially, the reasoning is the same as that presented for scenarios 1 and 2: when met with CD/DC plays, these emotional characters are able to exploit others before locking into CC plays. The safer exploitation of highly responsive emotional characters is admired more in scenario 3 since the likelihood of a CD/DC play being encountered initially is at its highest point (0.25, see table 7.4). This means that there is a greater likelihood that the over-exploitation gamble entailed by moderately responsive emotional characters will not pay off.

To emphasise how much the highly responsive emotional character group's safe exploitation is admired when CC is not the most likely initial play, consider the increased prevalence of highly responsive emotional characters compared to moderately responsive emotional characters. Highly responsive emotional characters have a 2/3 chance of establishing CC cycles from CD/DC plays and in scenario 3, it is more likely that CC is established initially by virtue of initial percentages of cooperators and defectors in the population. Therefore, this should entail a reduction in the prevalence of highly responsive emotional characters since the ability to convert CD/DC plays into CC is not as admirable given that initial cooperation is more likely. Despite this, highly responsive emotional characters are still more prevalent than moderately responsive emotional characters in scenario 3 and their prevalence increases from scenario 2 to 3.

Less responsive emotional characters are least prevalent in this scenario out of all scenarios considered due to the increased likelihood of CD/DC plays being established initially. Under such conditions, less responsive emotional characters have a 2/3 chance of establishing DD plays and only a 1/3 chance of maintaining stabilised/destabilised DC/CD plays. Therefore, the weakness of this emotional character is brought to light more than it is in other scenarios which results in its much reduced prevalence in this scenario.

Scenario 4 This scenario marks a turning point in trends of responsive emotional character prevalence as it does for tolerant emotional character prevalence in the same scenario. To reiterate, scenario 4 is the first scenario of scenarios 1-5 where initial defection rates are lower than initial cooperation rates and the probability of establishing CC initially is higher than the probability of establishing any other type of initial play (see table 7.4). This puts a greater focus upon the benefits conferred by the ability of moderately responsive emotional characters to over-exploit highly tolerant emotional characters since not many will perform such exploitation and there are many more targets to hit. Thus, an increase in the prevalence of moderately responsive emotional characters is observed.

The decrease in prevalence of highly responsive emotional characters is also explained by the increase in likelihood of CC being established initially. Since CC is the most likely

initial play in context of scenario 4, this means that the premium placed upon the benefits conferred by highly responsive emotional characters (2/3 chance of establishing CC from DC plays) is reduced. The increased premium upon over-exploitation by moderately responsive emotional characters also accounts for the reduced prevalence of highly responsive emotional characters because the benefits of safe exploitation are diluted as a consequence of the increased numbers of initial cooperators.

Finally, the increased prevalence of less responsive emotional characters does not appear to be that pronounced if scenarios 3 and 4 are compared. The increased prevalence of less responsive emotional characters may be simply due to the fact that agents with less responsive emotional characters were surrounded by highly tolerant emotional characters that initially cooperated (resulting in a reduced number of less responsive emotional characters establishing DD plays from CD/DC plays). The increase in prevalence observed is not so pronounced that any particular attribute being admired more in context of less responsive emotional characters can be ascribed, however.

Scenario 5 Scenario 5 is particularly interesting since it would appear that the ability to exploit opponents extensively, as conferred by moderately responsive emotional characters, is admired most prominently. In this scenario it is observed in table 7.12 and figure 7.2 that the prevalence of moderately tolerant emotional characters increases to such an extent that these emotional characters become more prevalent than highly responsive emotional characters.

The explanation underlying this observation is simply an extrapolation of that presented in scenario 4 above; because CC is so likely to be established initially between opponents, the ability of highly responsive emotional characters to convert 2/3 of CD/DC plays to CC plays is further devalued. In context of scenario 5, the ability to exploit others to earn individual scores greater than those obtained by cooperating with the opponent for every round of the game, is highly valued. Since moderately responsive emotional characters are capable of exploiting opponents for one round more than highly responsive emotional characters before locking into CC cycles, moderately responsive emotional characters are admired more.

The increase in prevalence of less responsive emotional characters is puzzling since one would expect their prevalence to decrease because they are not capable of exploiting opponents and they have a 2/3 chance of establishing DD cycles from CD/DC plays. The only explanation that can be given for this increase in prevalence is due to a spillover effect from selection of reduced tolerance in this scenario. Indeed, A1:G3 enjoys the largest increase in prevalence out of the less responsive emotional character groupings, as can be observed in table 7.16:

- A1:G3's frequency of placing (the less tolerant character of the less responsive character group) in positions 1-3 increases by 49 overall.
- A2:G3's frequency of placing (the moderately tolerant character of the less responsive character group) in positions 1-3 increases by 7 overall .

TABLE 7.16: Comparison of A1:G3, A2:G3 and A3:G3's frequency of placing in positions 1-3 for scenarios 4 and 5.

		Position		
	Scenario #	*1*	*2*	*3*
A1:G3	4	4	19	12
	5	23	30	31
A2:G3	4	5	6	12
	5	11	7	12
A3:G3	4	9	13	17
	5	7	14	19

- A3:G3's frequency of placing (the highly tolerant character of the less responsive character group) in positions 1-3 only increases by 1 overall.

7.5.2 Emotional Character Prevalence and Initial Impressionability Percentages

Headline Results

- As the initial percentage of less impressionable agents increases, highly tolerant and highly responsive emotional characters become prevalent due to their increased likelihood of establishing CC plays when CD/DC plays are encountered.

- Increasing the initial percentage of highly impressionable agents has no discernible effect since character prevalence is governed solely by probability in such circumstances. For example: if an emotional character is more concentrated in a neighbourhood, this emotional character is more likely to gain a foothold and propagate if there is a greater percentage of highly impressionable agents present in a population.

The discussion provided in this section is intended to answer research question 4b (see section 7.1). The research question itself is a little limited and requires some expansion in light of admiration being selected as the emotion used as a basis to allow agents to switch emotional character. The research question can now be phrased as "*what effect does altering the percentage of highly, moderately and less impressionable agents have upon emotional character prevalence within the population?*".

This question is addressed by scenarios 6-14 where the initial percentages of impressionable agent groups are the independent variables used (the percentages of initial cooperators and defectors remain constant at 50% in context of these scenarios, see table 7.2). Note that the scenarios themselves may be divided into three groups:

- Scenarios 6-8 focus upon the effects of increasing the percentage of highly impressionable agents in the population.

- Scenarios 9-11 focus upon the effects of increasing the percentage of moderately impressionable agents in the population.
- Scenarios 12-14 focus upon the effects of increasing the percentage of less impressionable agents in the population.

Admiration is elicited in highly impressionable agents when an agent in their comparator set achieves the highest individual score out of all other agents in this set (including the agent doing the comparing) once. Due to this, emotional characters are not under any great selection pressure in scenarios 6-8 since there is a high chance that there may be numerous emotional characters that an agent admires at the conclusion of the first five rounds of a game. This renders emotional character prevalence as a probability game: the greater the number of emotional characters in the comparator set that have elicited admiration in the admirer, the less chance that a particular emotional character has of being propagated. Exact odds are not provided since one would need to know the emotional characters of all agents in all comparator sets for every agent in the game and calculate, after every five rounds, what emotional character each agent admires. Instead, it is enough to say that emotional characters do not have to function very effectively to maintain and promote their prevalence; probability has a greater bearing upon emotional character prevalence in these scenarios.

In scenarios 9-11, where moderately impressionable agents are prevalent, the selection pressure upon emotional characters is increased since an emotional character must enable an agent to achieve the highest individual score of its comparator set twice before admiration is elicited. Subsequently, emotional character prevalence is less concerned with probability and more with the play styles that an emotional character bestows. The greatest selection pressure exerted upon emotional characters comes in context of scenarios 12-14 where less impressionable agents are most common. To elicit admiration in less impressionable agents, an emotional character must ensure that its agent attains the highest individual score three times. Therefore, if an emotional character becomes prevalent in these scenarios it is highly likely that this is because of its emotional character rather than mere chance.

Methodology To investigate this question, frequency of placing for all emotional characters over entire scenarios was calculated in the same way described in section 7.5.1. However, in this section frequencies of placing in positions 1-3 were calculated for scenarios 6-14 rather than scenarios 1-5.

Given that the particular play histories of tolerance and responsive groups were outlined in sections 7.5.1.1 and 7.5.1.2, this section simply comments upon why these plays are or are not selected for when the percentage of different admiration levels is altered in the population. As with the analysis of what effect altering the initial cooperation and defection rates has upon character prevalence (see section 7.5.1), this section will continue to consider emotional characters based upon tolerance/responsiveness groupings and accordingly this section will be further subdivided into two. Section 7.5.2.1

considers how varying levels of impressionability affect tolerance groups of emotional characters whilst section 7.5.2.2 assesses the effects in context of responsive character groups.

Tables 7.17 and 7.18 along with figures 7.3 and 7.4 show the frequency of placing in positions 1-3 for emotional character groupings based on tolerance (table 7.17 and figure 7.3) and responsiveness (table 7.18 and figure 7.4). To clarify the graphs presented: each major interval on the x-axis represents the frequency of placing in positions 1-3 for each level of impressionability of the three groups outlined earlier:

1. The first interval marks the frequency of placing in positions 1-3 for each tolerance or responsiveness group in context of scenarios 6, 9 and 12 (when the percentages of less, moderately and highly impressionable agents are equal to 50%).

2. The second interval marks the frequency of placing in positions 1-3 for each tolerance or responsiveness group in context of scenarios 7, 10 and 13 (when the percentages of less, moderately and highly impressionable agents are equal to 70%).

3. The third interval marks the frequency of placing in positions 1-3 for each tolerance or responsiveness group in context of scenarios 8, 11 and 14 (when the percentages of less, moderately and highly impressionable agents are equal to 90%)

These groupings are also clarified further by the style of the line used:

1. Solid lines indicate frequencies of placing in positions 1-3 for each tolerance or responsiveness group in context of scenarios 6, 9 and 12 (when the percentages of less, moderately and highly impressionable agents are equal to 50%).

2. Dotted lines indicate the frequency of placing in positions 1-3 for each tolerance or responsiveness group in context of scenarios 7, 10 and 13 (when the percentages of less, moderately and highly impressionable agents are equal to 70%).

3. Broken lines indicate the frequency of placing in positions 1-3 for each tolerance or responsiveness group in context of scenarios 8, 11 and 14 (when the percentages of less, moderately and highly impressionable agents are equal to 90%)

7.5.2.1 Effects Upon Tolerance

The basic prevalence of each tolerance group can be explained using the discussions contained in section 7.5.1.1: highly tolerant emotional characters have a 2/3 chance of establishing CC and a 1/3 chance of maintaining stabilised/destabilised CD/DC plays when met with CD/DC plays; moderately tolerant emotional characters have an equal chance of establishing CC/DD or maintaining stabilised/destabilised CD/DC plays following CD/DC plays; less tolerant emotional characters have a 2/3 chance of establishing DD from CD/DC plays and a 1/3 chance of maintaining stabilised/destabilised CD/DC

TABLE 7.17: Frequencies of placing in positions 1-3 for tolerance groups in scenarios 6-14.

	Scenario #	**A:1**	**A:2**	**A:3**
Ad:1	6	72	145	284
	7	79	151	267
	8	99	167	252
Ad:2	9	73	126	302
	10	65	135	297
	11	72	129	298
Ad:3	12	67	129	302
	13	61	135	297
	14	65	115	320

TABLE 7.18: Frequencies of placing in positions 1-3 for responsiveness groups in scenarios 6-14.

	Scenario #	**G:1**	**G:2**	**G:3**
Ad:1	6	239	172	90
	7	237	185	75
	8	232	194	82
Ad:2	9	241	179	81
	10	223	182	92
	11	236	169	94
Ad:3	12	207	191	100
	13	226	170	97
	14	230	183	87

plays when met with CD/DC plays. To facilitate a discussion of impressionability and its effects upon tolerance this section will be divided into three paragraphs for scenarios 6-8, 9-11 and 12-14 respectively.

High Impressionability Prevalent As stated in the introduction to section 7.5.2, when the population is comprised of highly impressionable agents, emotional character prevalence is governed more by probability than behavioural traits imparted by emotional characters. This is why the prevalence of highly tolerant emotional characters decreases, even though they have the highest likelihood of establishing CC plays from CD/DC plays. Their prevalence appears to decrease as the scenarios progress from 6 to 8 whilst the prevalence of less tolerant agents (who have the highest likelihood of establishing DD from CD/DC plays), increases.

Moderate Impressionability Prevalent When emotional character prevalence is determined more by the behaviour imparted by emotional character when an agent is faced with CD/DC plays, highly tolerant agents become more prevalent since their high chance (2/3) of converting CD/DC plays into CC is selected for. Thus an increase in prevalence both in context of impressionability groupings and in tolerance groupings is observed for highly tolerant emotional characters.

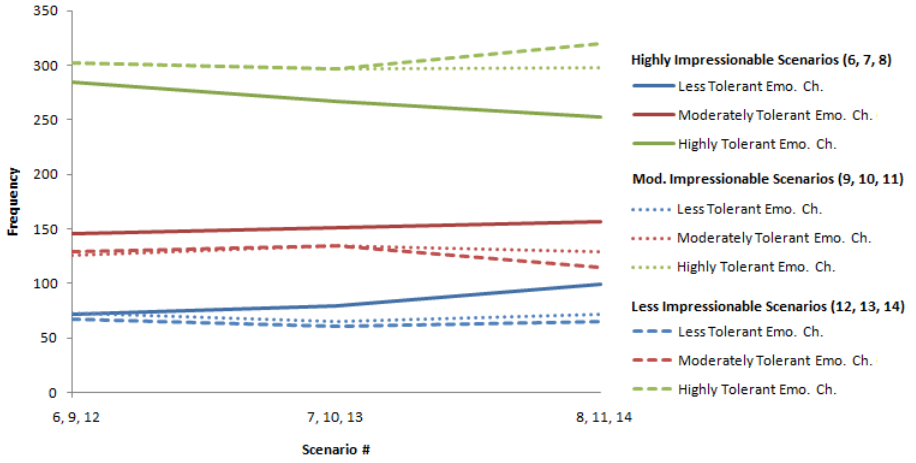

FIGURE 7.3: Frequencies of placing in positions 1-3 for tolerance groups in context of different levels of impressionability.

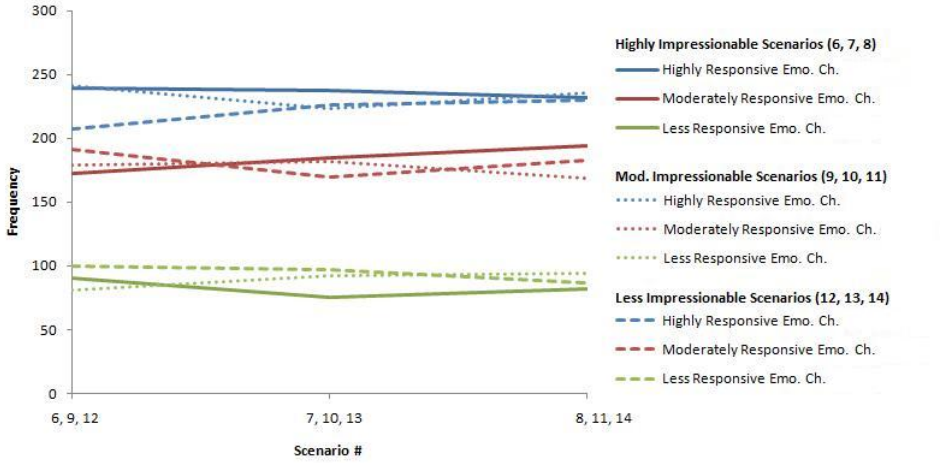

FIGURE 7.4: Frequencies of placing in positions 1-3 for responsiveness groups in context of different levels of impressionability.

Moderately and less tolerant agents are now selected against because their increased likelihood of establishing DD cycles from CD/DC plays is highlighted. In opposition to highly tolerant emotional characters, these emotional characters decrease in prevalence both in terms of impressionability groupings and tolerance groupings.

Low Impressionability Prevalent Under conditions where prevalence of emotional characters is determined mostly by the performance of agents in CD/DC plays by virtue of their emotional character, one can observe a focusing of emotional character prevalence.

With regards to highly tolerant emotional characters, their prevalence would appear to generally increase as more and more less impressionable agents are introduced into the population. Clearly, under situations where CC, CD/DC and DD plays all have an equal chance of occurring, high tolerance is greatly admired. This assertion is supported

by considering the results obtained for moderately and less tolerant agents in the same scenarios.

The prevalence of moderately tolerant agents noticeably decreases as greater selection pressure is applied to the emotional characters. This is because the increase in prevalence of highly tolerant characters must come from somewhere and for agents who are moderately tolerant initially, it is better to become highly tolerant since they will prove themselves more successful as the percentage of less impressionable agents increases. The prevalence of less tolerant characters remains relatively static since there are some benefits to not being exploited too readily in initial rounds. However, after DD plays are locked into by all neighbours there is not a substantial chance that these agents will change their emotional characters.

7.5.2.2 Effects Upon Responsiveness

Like section 7.5.2.1 above, the basic prevalence of each responsiveness group can be explained by outlining how different responsiveness groups deal with DC/CD plays (discussed in section 7.5.1.2). Highly responsive emotional characters have a 2/3 chance of establishing CC and a 1/3 chance of maintaining stabilised/destabilised CD plays when met with DC/CD plays; moderately responsive emotional characters have an equal chance of establishing CC/DD or maintaining stabilised/destabilised CD/DC plays following DC/CD plays; less responsive emotional characters have a 2/3 chance of establishing DD and a 1/3 chance of maintaining stabilised/destabilised CD/DC plays when met with DC/CD plays.

As discussed previously in section 7.5.1.2: moderately responsive emotional characters are also able to exploit highly tolerant opponents for one round more than highly responsive emotional characters whilst still being able to establish CC afterwards. Less responsive emotional characters on the other hand attempt to exploit moderately tolerant opponents too much resulting in DD being established.

To facilitate a discussion of impressionability and its effects upon relevance this section will also be divided into three paragraphs for scenarios 6-8, 9-11 and 12-14 respectively.

High Impressionability Prevalent When impressionability is high, emotional character prevalence is driven more by chance (see section (see section 7.5.2.1)) i.e. if an emotional character is competing against seven others it has less of a chance of being adopted by an admirer than it would if all other agents in the same comparator set were of the same emotional character. Consequently, discussion of the prevalence patterns for these scenarios is not provided.

Moderate Impressionability Prevalent When emotional characters have to be successful over a greater period of time to be admired, it can be observed that, whilst the prevalence of highly responsive emotional characters falls as the scenario changes

from 9 to 10, moderately responsive emotional characters become more prevalent. This is for two reasons: firstly, the two round exploitation made possible by moderately responsive agents is selected for since it gives a higher individual score than the single round exploitation achieved by highly responsive agents (see section 7.5.1.2). Secondly, highly impressionable agents are still numerous in the population and the increase in prevalence of moderately responsive emotional characters may be due to this. This line of reasoning is justified by the prevalence trends that are observed as the scenario number progresses from 8 to 9.

There is also an observed increase in prevalence of less responsive and highly responsive emotional characters from scenario 9 to 11 and a decrease in prevalence of moderately responsive emotional characters. The explanation for this is similar to that provided above i.e. it may be mostly due to chance with respect to initial comparator sets (there may be a greater number of less responsive emotional characters in certain comparator sets so their prevalence is increased due to their concentration). No other reasonable explanation can be proposed for this observation.

Low Impressionability Prevalent Generally speaking, when the majority of the population are less impressionable, highly responsive emotional characters are admired more and less responsive emotional characters are admired less. This is due to the reasons already explained i.e. the increased probability of highly responsive emotional characters establishing CC and less responsive emotional characters establishing DD from DC/CD plays.

What is interesting is the apparent dip in prevalence of moderately responsive emotional characters as the scenario number is increased from 12 to 13. This is most likely due to the over-exploitive nature of moderately responsive emotional characters being admired less when emotional characters must promote individual scores for a longer period of time. However, the observed increase in prevalence of moderately responsive emotional characters in scenario 14 confounds the issue further. One can only suggest that the dip in prevalence observed in scenario 13 is because safe exploitation is again admired but then in scenario 14 both safe and unsafe exploitation are valued (unsafe more than safe, however).

7.5.3 Emotional Character Prevalence and Player/Comparator Sets

Headline Results

- Increasing the number of players in player and comparator sets exaggerates and concentrates emotional character prevalence. Increasing the number of agents in the player set exerts a greater effect in this respect, however.
- Highly tolerant emotional characters are admired most and tolerance itself is admired more than responsiveness. This is justified by the observation that less

responsive emotional characters in the same tolerance group were admired less before reduced tolerance.

In considering the effects of altering player/comparator sets upon emotional character prevalence I would posit that, by increasing the number of agents in these sets, emotional characters are subjected to greater selection pressures. The reasoning for this is such: if agents have to play against more opponents initially then the variety of emotional characters they play against is increased. Therefore, only emotional characters that perform well against a variety of others will prove to be successful and will propagate themselves throughout the population. By increasing the number of agents that comprise the comparator set then the same is also true except that emotional characters will have greater competition from others with respect to their effect upon behaviour. The intention of this section is to investigate the correctness of this hypothesis i.e. is it the case that increasing the number of agents in the player and comparator sets causes emotional character prevalence to become more pronounced and, if so, which emotional characters become more prevalent?

Methodology To investigate this question I have calculated the frequency that each emotional character places in positions 1-3 in individual sub-scenarios in context of all major scenarios. These calculated values are not listed here since there are so many but comparisons can be made between the effects of player set and comparator set upon tolerance and responsiveness groupings using these values. Due to this, this section will be split into two sub-sections: sub-section 7.5.3.1 will discuss the effects upon tolerance and responsiveness emotional character groupings when the player set is the independent variable. Alternatively, sub-section 7.5.3.2 considers the effects upon tolerance and responsiveness emotional character groupings when the comparator set is the independent variable in the simulation.

7.5.3.1 Effects of Player Set Upon Emotional Character Prevalence

In this section I will consider both general trends exerted by player set alterations with respect to the frequency of placing in positions 1-3 for each emotional character and whether or not increasing the number of agents in the player set *exaggerates* placement frequencies for emotional characters and whether frequencies of placement become more *concentrated* for each emotional character as player set members are increased.

General Trends To analyse the general effects of increasing the number of agents in the player set upon emotional character prevalence, four groups of sub-scenarios are considered in turn. These sub-scenario groupings ensure that the comparator set remains constant but the player set increases as sub-scenario number increases in a group. The groupings are listed below for clarification:

1. Group 1: Consists of sub-scenarios 1, 5, 9, 13 (comparator set is always *parallel* i.e. two agents make up this set).

2. Group 2: Consists of sub-scenarios 2, 6, 10, 14 (comparator set is always *orthogonal* i.e. four agents make up this set).

3. Group 3: Consists of sub-scenarios 3, 7, 11, 15 (comparator set is always *diagonal* i.e. four agents make up this set).

4. Group 4: Consists of sub-scenarios 4, 8, 12, 16 (comparator set is always *octagonal* i.e. eight agents make up this set).

If the relevant frequencies of placing for these four sub-scenario groups are plotted, four graphs are produced: these are illustrated in figures 7.5, 7.6, 7.7 and 7.8; it is from these graphs that the following analysis is constructed.

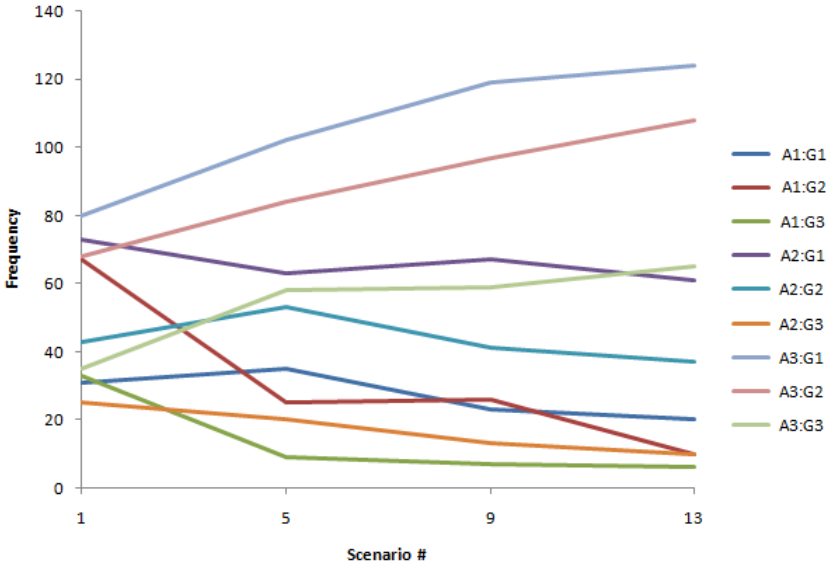

FIGURE 7.5: Frequency of placing in positions 1-3 for all emotional characters in sub-scenario group 1.

Considering the results in figures 7.5, 7.6, 7.7 and 7.8 it would appear that increasing the number of agents in the player set does not exert any discernible effect upon the ordering of emotional character prevalence. Without exception, the ordering of emotional character prevalence always resolves to A3:G1, A3:G2, A3:G3, A2:G1 and A2:G2 whilst emotional characters A1:G1, A1:G2, A1:G3 and A2:G3 are always least prevalent but their ordering follows no consistent pattern. It can therefore be asserted that, as the potential variety of emotional characters faced by an emotional character increases, it is the *tolerance* of emotional characters that is admired, irrespective of comparator set configuration. The results make this abundantly clear since highly tolerant emotional characters place more frequently in positions 1-3 as the number of agents in the player set increases and emotional characters that are less responsive but equally tolerant are admired more than less tolerant characters who are equally as responsive.

The assertion that tolerance is admired more than responsiveness is further fortified by considering the gradients of trend-lines that may be plotted for each of the emotional

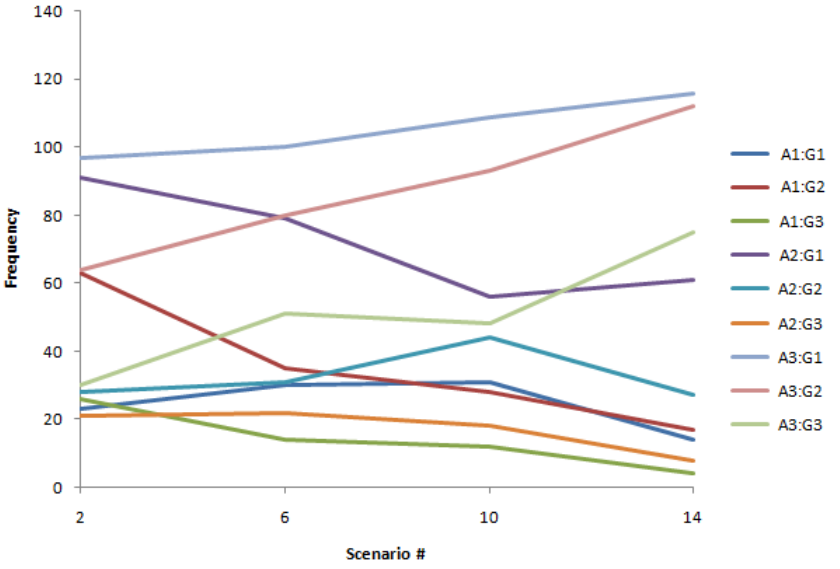

FIGURE 7.6: Frequency of placing in positions 1-3 for all emotional characters in sub-scenario group 2.

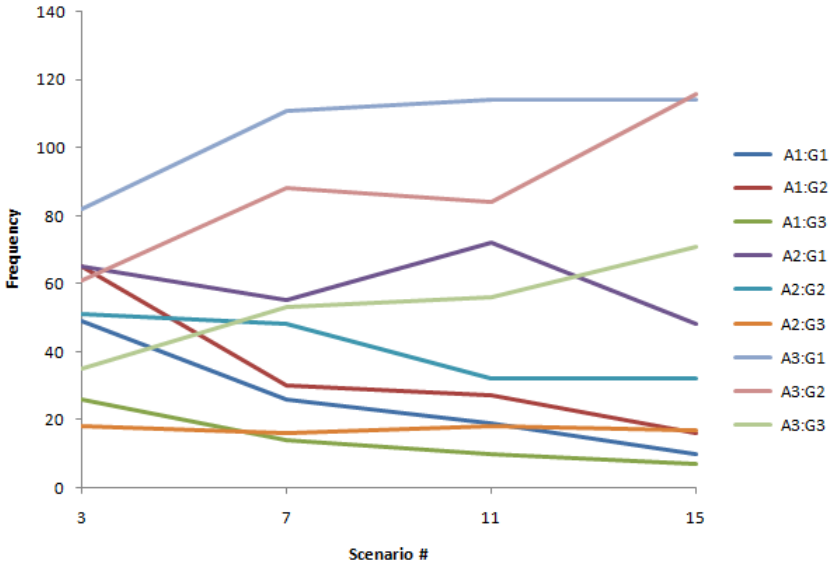

FIGURE 7.7: Frequency of placing in positions 1-3 for all emotional characters in sub-scenario group 3.

characters in figures 7.5, 7.6, 7.7 and 7.8. These trend-line gradients are presented in table 7.19. Again, without exception, highly tolerant emotional characters place more frequently in positions 1-3 irrespective of comparator set as the number of agents in the player set is increased. This is demonstrated by the observation that, in table 7.19, the only positive trend-line gradients are attributed to the highly responsive emotional characters (A3:G1, A3:G2 and A3:G3). All other emotional characters either suffer a decrease in frequency of placing in positions 1-3 or their increase and decrease of frequency as player set members are increased, cancels out to zero. With respect to

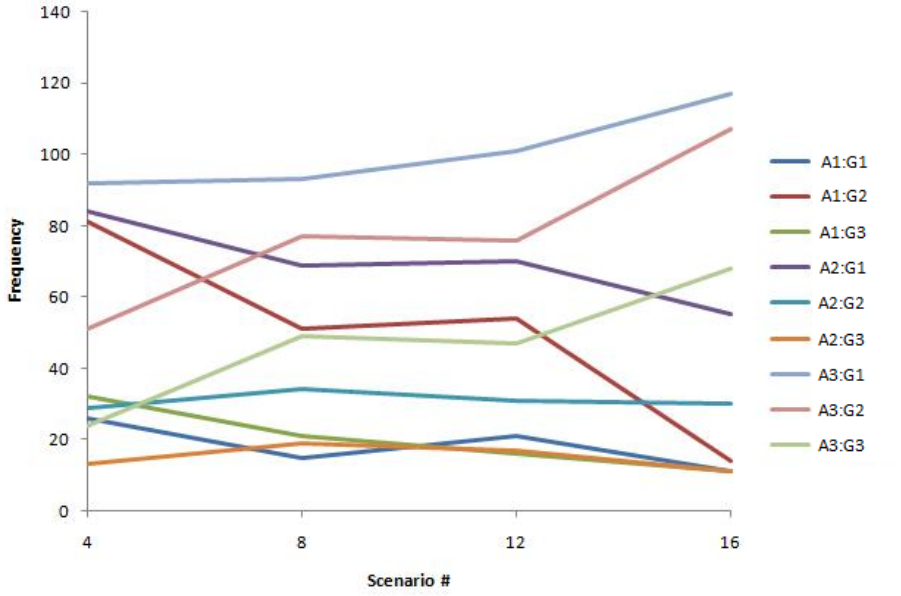

FIGURE 7.8: Frequency of placing in positions 1-3 for all emotional characters in sub-scenario group 4.

TABLE 7.19: Trend-line gradients of frequency of placing in positions 1-3 for all emotional characters in context of sub-scenario groups 1-4.

	Trend-line Gradients For Sub-Scenario Group n			
Emo. Ch.	$n = 1$	$n = 2$	$n = 3$	$n = 4$
A1:G1	-4.5	-2.6	-12.4	-3.9
A1:G2	-17	-14.5	-15	-19.8
A1:G3	-8.3	-6.8	-6.1	-6.8
A2:G1	-3.2	-11.3	-3.4	-8.6
A2:G2	-3	0	-7.3	0
A2:G3	-5.2	-4.3	-0.1	-0.8
A3:G1	14.9	6.6	9.9	8.3
A3:G2	13.3	15.7	16.1	16.7
A3:G3	9.1	13.2	11.1	13

these figures there are no trends that may be identified when emotional characters are considered in terms of their responsiveness.

Exaggeration and Concentration Trends To investigate whether increasing the number of agents in the player set causes an exaggeration and concentration of character prevalence frequencies, I have calculated both the average frequency that each emotional character places in positions 1-3 and the standard deviations of these placement frequencies. By calculating average frequencies of placing for emotional characters I intend to comment upon whether emotional character prevalence trends that have been identified in sections 7.5.1 and 7.5.2 are exaggerated by increasing the number of agents present in all player sets. Standard deviations of these placing frequencies will be used to determine

if increasing the agents in the player set causes a concentration of these frequencies for emotional characters.

In this section I will make use of a different set of sub-scenario groupings from those used previously. These alternative sub-scenario groupings are used as I am looking to establish whether emotional character prevalence trends change as *player set* configurations are altered. Thus, instead of comparing sub-scenario groups where player set configurations change homogeneously, this section compares sub-scenario groups where the player set is heterogeneous. For example: if I were to calculate and compare the average frequencies of placing for A1:G1 agents in context of the sub-scenario groupings previously used then, I will be comparing trends between sub-scenario groups 1 and 2 for example. In such a case, any trends that emerge are likely to be the result of the comparator set configuration since this is the variable that differs between sub-scenarios. Whilst player set configurations do increase in context of individual sub-scenario groups, they differ homogeneously between different sub-scenario groups. The sub-scenario groupings used in this section are listed below for clarification and reference:

1. Group 5: Consists of sub-scenarios 1, 2, 3, 4 (player set is always *parallel*).
2. Group 6: Consists of sub-scenarios 5, 6, 7, 8 (player set is always *orthogonal*).
3. Group 7: Consists of sub-scenarios 9, 10, 11, 12 (player set is always *diagonal*).
4. Group 8: Consists of sub-scenarios 13, 14, 15, 16 (player set is always *octagonal*).

If increasing the number of agents in the player set does indeed exaggerate general character prevalence trends then I would expect average frequency values to increase as the sub-scenario group considered increases from 5 to 8. In contrast, if it is true that increasing the number of agents in the player set increases the concentration of placing frequencies, I would expect standard deviation values to decrease as the sub-scenario group number increases. Plotting the frequencies of placing for all emotional characters in sub-scenario groups 5-8 results in figures 7.9, 7.10, 7.11 and 7.12.

A few interesting observations emerge from comparing the results displayed in figures 7.9, 7.10, 7.11 and 7.12. Firstly, it would appear that, by increasing the number of agents in the player set, familiar character prevalence patterns develop. For example: if the frequencies of placing in positions 1-3 for all emotional characters in sub-scenario group 5 (see figure 7.9) are visually compared against the frequencies of placing in positions 1-3 for all emotional characters in sub-scenario group 8 (see figure 7.12), there is a clear separation and refinement of emotional character prevalence. Indeed, the character prevalence patterns noted by a consideration of table 7.19 are emphasised as the number of agents in the comparator set is increased.

The average frequency by which each emotional character places in positions 1-3 in context of sub-scenarios 5-8 and standard deviation of these values can be used to fortify these assertions regarding emergence of character prevalence trends identified earlier (see table 7.19). The raw values derived are presented graphically in figures 7.13 and 7.14.

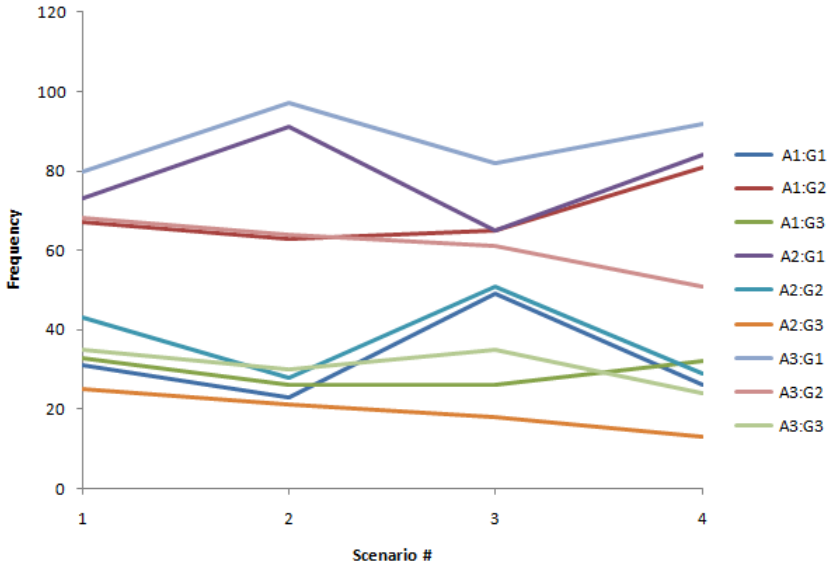

FIGURE 7.9: Frequency of placing in positions 1-3 for all emotional characters in sub-scenario group 5.

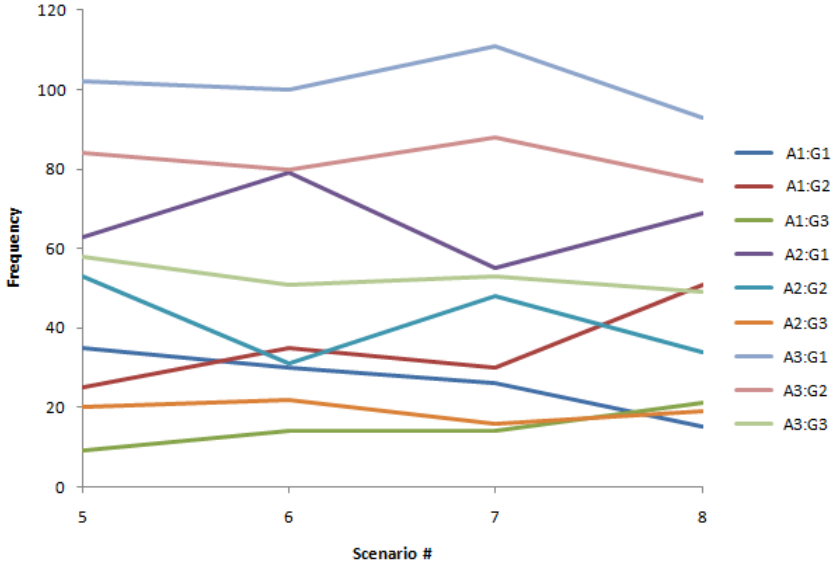

FIGURE 7.10: Frequency of placing in positions 1-3 for all emotional characters in sub-scenario group 6.

If trend-line gradients are then calculated for each emotional character in context of figures 7.13 and 7.14 then general trends present in the data can be identified without going into unnecessary detail. Rather than facilitating explanation of trends, such detail may only confound explanations. The trend-line gradients derived are listed in table 7.20.

It is clear from an analysis of the data contained in table 7.20 that increasing the number of agents in the player set both exaggerates already noted character prevalence trends and concentrates prevalence frequencies. As can be seen in table 7.20 and figure

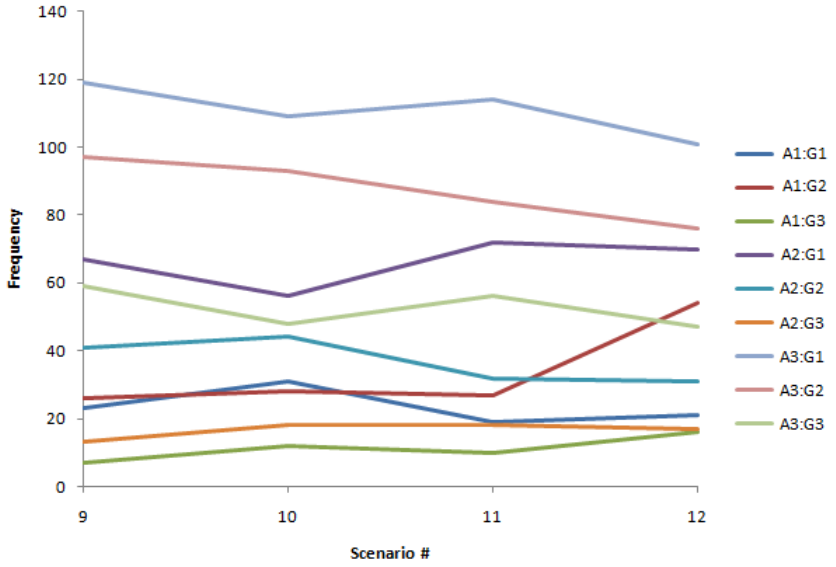

FIGURE 7.11: Frequency of placing in positions 1-3 for all emotional characters in sub-scenario group 7.

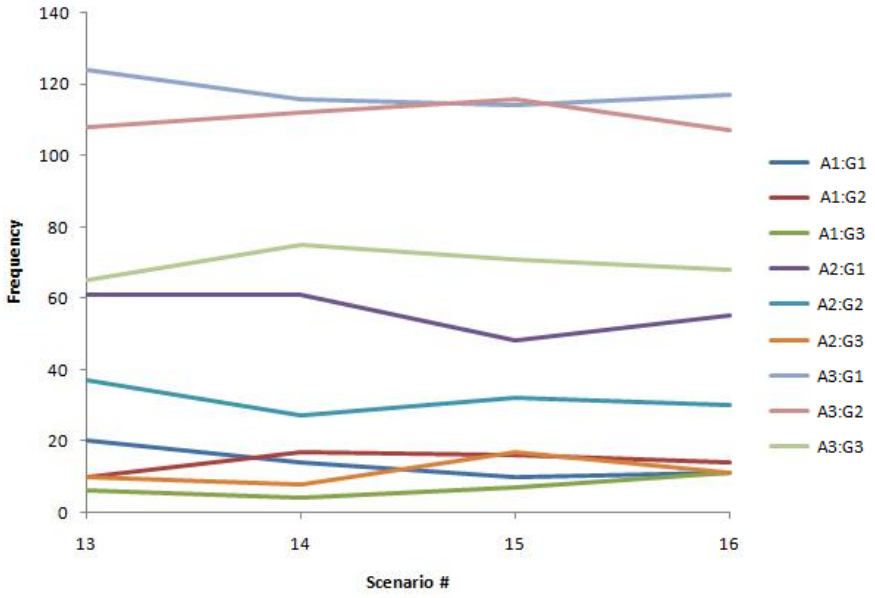

FIGURE 7.12: Frequency of placing in positions 1-3 for all emotional characters in sub-scenario group 8.

7.13, every emotional character except A3:G1, A3:G2 and A3:G3 experiences an overall decrease in its average prevalence in positions 1-3 as the number of agents comprising player sets for agents increases. Conversely, emotional characters A3:G1, A3:G2 and A3:G3 are the only emotional characters that experience any form of increase in prevalence. This would again indicate that, as emotional characters are forced to perform better against a wide variety of opponents, it is tolerance that is most valued along with high responsiveness.

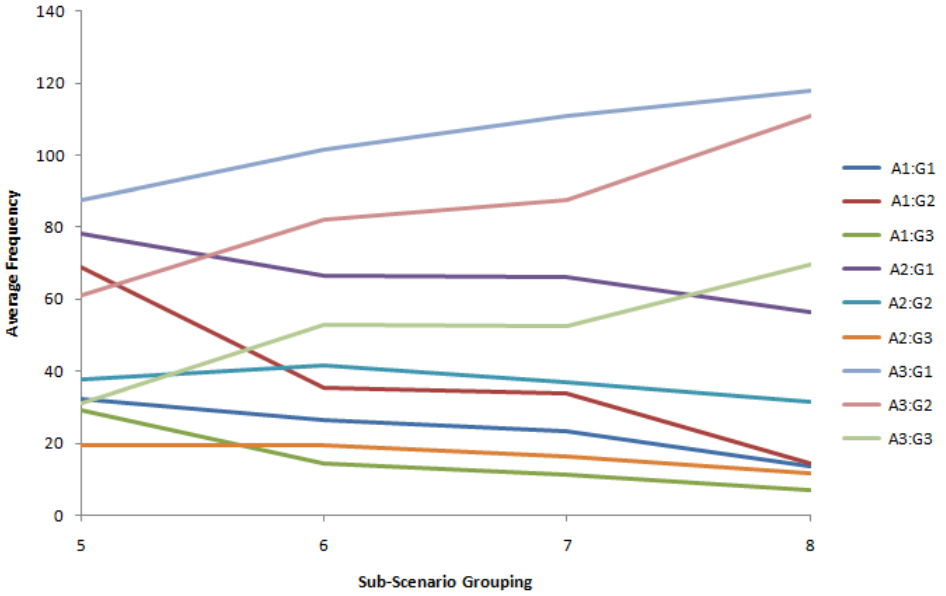

FIGURE 7.13: Average frequency of placing in positions 1-3 for each emotional character in context of sub-scenario groups 5-8.

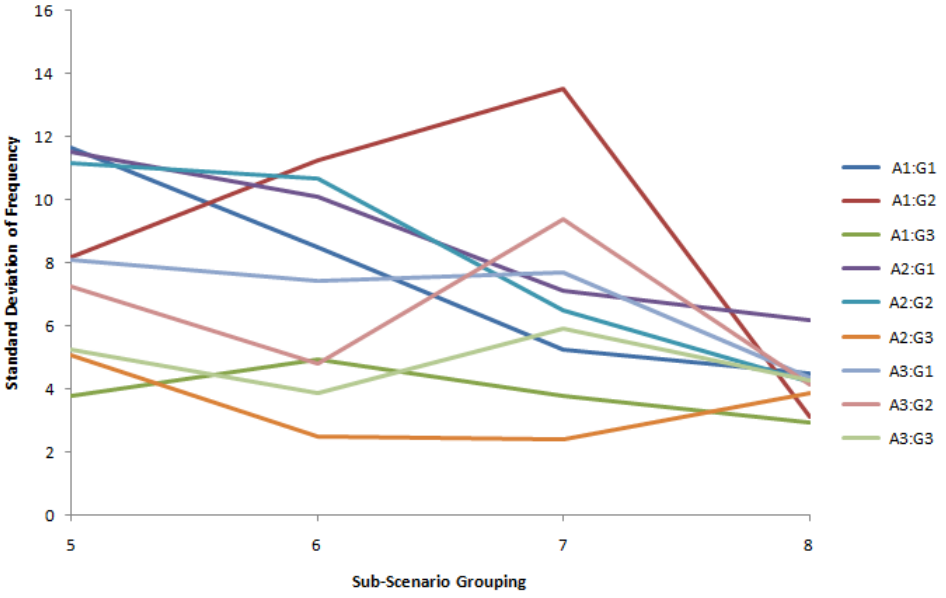

FIGURE 7.14: Standard deviations of placing in positions 1-3 for each emotional character in context of sub-scenario groups 5-8.

Furthermore, every emotional character's standard deviation of placing frequency is subjected to a general reduction as numbers of agents in the player set increases (see table 7.20 and figure 7.14). This ultimately demonstrates that emotional character prevalence is concentrated as the number of agents in the player set increases. Therefore, if others are interested in ascertaining whether particular emotional characters perform better given different initial conditions then, by increasing the variety of emotional characters that an agent must play against, selection pressures upon characteristics are increased.

TABLE 7.20: Trend-line gradients for average and standard deviations of placing frequency in positions 1-3 in context of sub-scenario groups 5-8.

Emo. Ch.	**Average Frequency Trend-line Slope**	**Standard Deviation Trend-line Slope**
A1:G1	-5.85	-2.47
A1:G2	-16.57	-1.29
A1:G3	-7	-0.37
A2:G1	-6.63	-1.9
A2:G2	-2.33	-2.47
A2:G3	-2.6	-0.37
A3:G1	9.93	-1.1
A3:G2	15.5	-0.48
A3:G3	11.6	-0.08

Consequently, behaviour beneficial to maximisation of individual score conferred by emotional characters becomes more prevalent throughout the population and behaviour detrimental to maximisation of individual score becomes less prevalent.

Given these observations, it would appear that high tolerance and responsiveness are admirable characteristics yet it is tolerance that is admired more than responsiveness. This is particularly interesting since tolerance is more concerned with establishment of CC plays whereas responsiveness is concerned with exploitation of opponents (see section 7.5.1). If success of an emotional character is determined by individual score and it is tolerance that is admired more than responsiveness then this would appear to indicate that increased self-interest is *self-defeating* when random defection and cooperation is not possible within a population.

7.5.3.2 Effects of Comparator Set Upon Emotional Character Prevalence

This section will be structured in much the same way as section 7.5.3.1 i.e. I will again consider general trends exerted upon emotional character prevalence and whether or not placement frequencies for emotional characters are *exaggerated* and *concentrated* except in this section the effect of increasing the number of agents in the comparator set upon these factors is investigated.

General Trends To determine if any general trends in character prevalence hold as the number of agents in the comparator set increases, irrespective of player set configurations, I will analyse the frequency of placing in positions 1-3 for each emotional character in context of sub-scenarios 5-8. Graphs containing the relevant values have already been produced (see figures 7.9, 7.10, 7.11 and 7.12). The sub-scenario groups used in this section are listed again below for ease of reference:

1. Group 5: Consists of sub-scenarios 1, 2, 3, 4 (player set is always *parallel* i.e. two agents make up this set).

TABLE 7.21: Trend-line gradients of frequency of placing in positions 1-3 for all emotional characters in context of sub-scenario groups 5-8.

	Trend-line Gradients For Sub-Scenario Group n			
Emo. Ch.	$n = 5$	$n = 6$	$n = 7$	$n = 8$
A1:G1	-1.9	-6.4	-1.8	-3.1
A1:G2	4.4	7.3	8.3	1.1
A1:G3	-0.3	3.6	2.5	1.8
A2:G1	0.7	-0.6	2.5	-3.1
A2:G2	-1.9	-4	-4.2	-1.6
A2:G3	-3.9	-0.9	1.2	1.2
A3:G1	2.1	-1.6	-4.9	-2.3
A3:G2	-5.4	-1.3	-7.2	0.1
A3:G3	-2.8	-2.5	-2.8	0.5

2. Group 6: Consists of sub-scenarios 5, 6, 7, 8 (player set is always *orthogonal* i.e. four agents make up this set).

3. Group 7: Consists of sub-scenarios 9, 10, 11, 12 (player set is always *diagonal* i.e. four agents make up this set).

4. Group 8: Consists of sub-scenarios 13, 14, 15, 16 (player set is always *octagonal* i.e. eight agents make up this set).

Unlike the comparisons of general character prevalence trends between sub-scenario groups 1, 2, 3 and 4 (see section 7.5.3.1), trends that emerge from such a comparison between sub-scenario groups 5, 6, 7 and 8 are less clear. Even plotting trend-lines and analysing their gradients does not yield any clearer insights (see table 7.21). One of the most notable observations however is that A3:G1 is always most prevalent in positions 1-3 no matter what sub-scenario group is considered out of 5, 6, 7 and 8. Similarly, A3:G2 is also prevalent in positions 2 or 3 but its prevalence is not as consistent as it is when comparing general trends in emotional character prevalence in context of sub-scenario groups 1-4.

If sub-scenario group 8 is considered the results observed fall back into the familiar tolerance divisions that were discussed for emotional characters when considering general trends observed when the player set is altered (see section 7.5.3.1): highly tolerant characters place most frequently in positions 1-3 followed by moderately tolerant characters and then less tolerant characters. In addition, responsiveness divisions are also maintained: less responsive emotional characters in the same tolerance band place less frequently in positions 1-3 but still more so than emotional characters who are less tolerant.

Exaggeration and Concentration Trends To determine whether increasing the number of agents in the comparator set exaggerates and concentrates placing frequencies for emotional characters in positions 1-3, sub-scenario groups 1-4 will be considered.

These sub-scenario groups are used as I am now looking to consider trends in average placement frequencies and standard deviations in context of comparator set differences. These sub-scenarios are listed again below for ease of reference:

1. Group 1: Consists of sub-scenarios 1, 5, 9, 13 (comparator set is always *parallel* i.e. two agents make up this set).
2. Group 2: Consists of sub-scenarios 2, 6, 10, 14 (comparator set is always *orthogonal* i.e. four agents make up this set).
3. Group 3: Consists of sub-scenarios 3, 7, 11, 15 (comparator set is always *diagonal* i.e. four agents make up this set).
4. Group 4: Consists of sub-scenarios 4, 8, 12, 16 (comparator set is always *octagonal* i.e. eight agents make up this set).

Average frequency of placing in positions 1-3 and associated standard deviation values for each emotional character is context of sub-scenario groups 1-4 are presented graphically in figures 7.15 and 7.16.

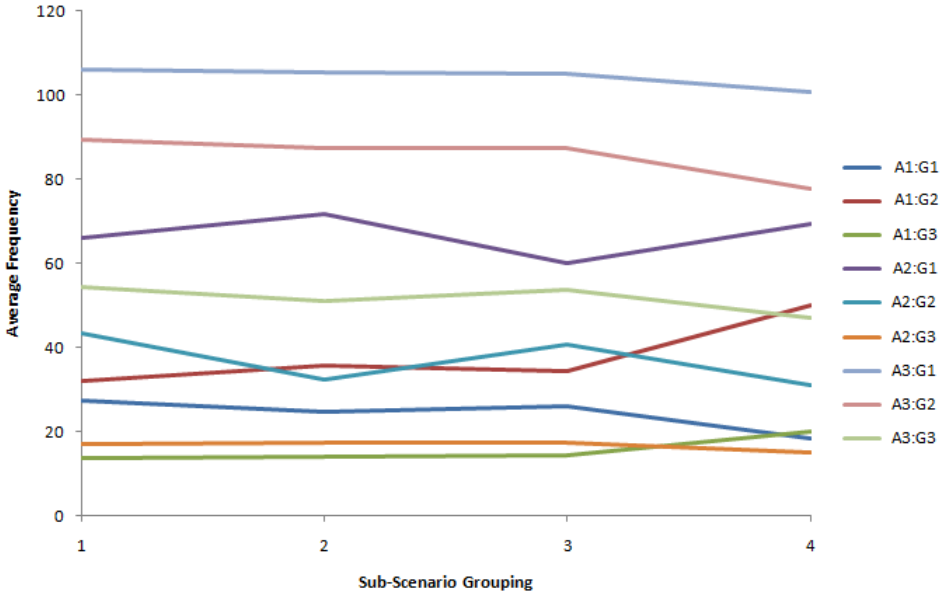

FIGURE 7.15: Average frequency of placing in positions 1-3 for each emotional character in context of sub-scenario groups 1-4.

With respect to figure 7.15 it would appear that increasing the number of agents in the comparator set exerts little effect upon average frequencies of placing in positions 1-3 for emotional characters. There also appears to be no strong effect either way upon standard deviations of frequencies for placing in positions 1-3 for emotional characters (see figure 7.16). Plotting general trend-lines and deriving their gradients for average frequencies of placing in positions 1-3 and associated standard deviations in context of sub-scenario groups 1-4 also gleans little information (see table 7.22).

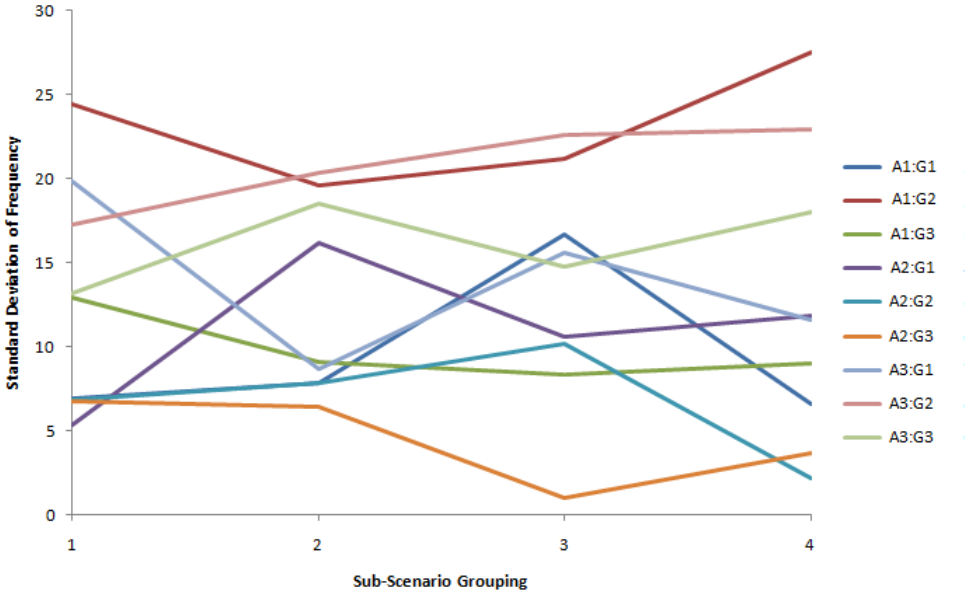

FIGURE 7.16: Standard deviations of placing in positions 1-3 for each emotional character in context of sub-scenario groups 1-4.

TABLE 7.22: Trend-line gradients for average and standard deviations of placing frequency in positions 1-3 in context of sub-scenario groups 1-4.

Emo. Ch.	**Average Frequency Trend-line Slope**	**Standard Deviation Trend-line Slope**
A1:G1	-2.55	0.8
A1:G2	5.3	1
A1:G3	1.9	-1.2
A2:G1	-0.1	1.4
A2:G2	-2.9	-1.2
A2:G3	-0.6	-1.5
A3:G1	-1.7	-1.8
A3:G2	-3.5	1.9
A3:G3	-1.9	1.1

Considering the data given in table 7.22, it would appear that, by increasing the number of agents in the comparator set, placing frequency trends are not positively exaggerated as they are when the number of agents in the player set are altered (see table 7.20). In fact, the opposite situation occurs in the majority of cases i.e. the average frequency of placing in positions 1-3 for emotional characters decreases as the number of agents in the comparator set increases. This inverse relationship is not uniform (see relevant values for A1:G2 and A1:G3), but it does exist nonetheless. Such a result is somewhat unexpected but understandable since increasing the number of emotional characters in the comparator set would increase the number of emotional characters that may be potentially admired thus making it harder for an emotional character to

be admired. Therefore, by increasing the number of agents in the comparator set, the more diluted character prevalence becomes.

In terms of standard deviations of placing in positions 1-3, the more agents in the comparator set, the less concentrated prevalence frequency becomes. However, like the average frequency values discussed above, this trend is not uniform across all emotional characters since the standard deviations for emotional characters A1:G3, A2:G2, A2:G3 and A3:G1 decreases as the number of agents in the comparator set increases. The reasoning for this is similar to when average frequencies of placing in positions 1-3 were considered: when emotional characters are more likely to be compared against a greater number of other emotional characters, the benefits conferred by their characteristics are diluted.

By considering this information it would not be unreasonable to assert that player set alterations give a better indication of which characteristics conferred by an emotional character are beneficial. Altering comparator set configurations seems to dilute these effects but in a very weak fashion so no meaningful information can be gleaned from an analysis of prevalence frequencies using differing comparator sets.

7.5.4 Overall Emotional Character Prevalence

The intention of this final section is to ascertain whether there exists an emotional character in the simulations that is more prevalent than others irrespective of initial conditions. Determining the answer to this question simply involves summing together the frequencies in which each character placed in each position in each scenario. For example: if A1:G1 placed first 4 times overall in each of the 14 scenarios then its total frequency of placing in first place overall would be 4×14. These total frequencies of placing in first position for each emotional character over all scenarios are shown in table 7.23. The maximum frequency value obtained for each position is highlighted in bold. Note that little weight should be placed upon middle placings due to the small variations in numbers: the point made by this table is that emotional characters whose anger activation threshold is greater than their gratitude activation threshold are very successful.

With respect to first place, A3:G1 is most prevalent placing first 951 times over all scenarios. The reason for the success of this emotional character is as described in section 6.5 of chapter 6 and as further elaborated in sections 7.5.1 and 7.5.2 in this chapter. Essentially, due to A3:G1's high tolerance and responsiveness, it has the greatest possible chance of establishing CC plays irrespective of its initial behaviour in CD/DC plays. Furthermore, A3:G1 avoids the establishment of DD plays from CD/DC plays aiding its maximisation of individual score.

What is especially interesting is that A3:G1 maximises total system score values to a greater extent than other emotional characters (as deduced and argued for in chapter 6) yet the emotional character is admired on the basis of how well it promotes *individual score*. Therefore, admiration would appear to give rise to an *emergent* property: even

TABLE 7.23: Total frequency of placing first in all scenarios for each emotional character.

	Position Frequency								
Emo. Character	*1st*	*2nd*	*3rd*	*4th*	*5th*	*6th*	*7th*	*8th*	*9th*
A1:G1	72	123	189	234	338	378	384	352	170
A1:G2	215	185	209	252	274	301	336	301	167
A1:G3	45	93	110	171	226	281	386	**509**	**419**
A2:G1	305	353	**411**	**373**	280	208	150	118	42
A2:G2	98	202	291	321	**369**	332	293	237	97
A2:G3	41	91	134	238	281	**404**	**447**	386	218
A3:G1	**951**	451	269	214	140	91	70	37	17
A3:G2	415	**562**	389	264	217	161	107	88	37
A3:G3	157	278	389	370	284	268	239	170	85
Max Value	951	562	411	373	369	404	447	509	419

though agents admire others on the basis of their individual score they are, unknowingly, promoting the acquisition of greater total system scores. The exact extent to which total system scores are promoted is not determinable in these simulations due to the fact that no control group consisting of agents with no emotional character, was implemented.

Analysing the emotional characters that have placed in first, second and third positions also provides interesting observations. As can be seen in table 7.23, A3:G2 places second most frequently over all scenarios whilst A2:G1 places third most frequently over all scenarios. This would indicate that reduced responsiveness is admired more than reduced tolerance since A3:G2 is just as tolerant as A3:G1 but is less responsive. A2:G1 on the other hand is just as responsive as A3:G1 but is less tolerant. This is most likely due to the fact that the reduced responsiveness embodied by A3:G2 benefits an agent since it is capable of increasing its individual score by exploiting an opponent for up to two rounds. Such a benefit is admired even if there is a risk of reduced individual score compared to the safer exploitation of A3:G1 (see section 7.5.1.2). The reduced tolerance of A2:G1 on the other hand only has a benefit in that it guards against the agent being exploited. The benefit conferred still inflicts a loss on the agent and it is just the extent of that loss which is moderated by reduced tolerance.

The pattern observed above also occurs when the characters who place most frequently in positions four to six and seven to nine are considered. With regards to positions four to six, A2:G1 again places most frequently in position four and the less responsive A2:G2 places in fifth position. Furthermore, the least responsive character of this tolerance group, A2:G3, places in sixth position. A2:G3 retains prevalence in seventh position overall but the less tolerant A1:G3 places most frequently in eighth and ninth positions. A1:G3 places in the lowest two positions due to its reduced responsiveness which gives it a 2/3 chance to establish DD cycles when an initially defecting agent of this emotional character meets an initially cooperative opponent.

7.6 Chapter Summary

In this chapter it was my intention to develop a computational mechanism inspired by human emotion to facilitate the propagation of the most successful basic emotional characters outlined in chapter 6 throughout the test-bed population in context of an iterated Prisoner's Dilemma game. The ultimate intention of developing such a mechanism was to investigate what effect various initial conditions had upon emotional character prevalence and whether or not one emotional character becomes prevalent in a population irrespective of the initial conditions used and why. These intentions led to the identification of admiration as the emotion to be used as a basis for the computational mechanism implemented following a consideration of literature that is concerned with this emotion's effect upon people. Using this literature I was also able to specify the eliciting conditions for this emotion in the context of human beings and from here translated these effects and eliciting conditions into computational mechanisms able to be implemented and used by agents within a large-scale MAS simulation.

To investigate the effects of differing initial conditions upon character prevalence I developed a new simulation environment based upon the simulation environment created in chapter 6. This facilitated the specification of initial percentages of cooperators and defectors, initial percentages of differing admiration activation thresholds and different configurations of player and comparator sets. The effect of varying these initial conditions upon emotional character prevalence was then studied so as to provide novel insights that have as yet been unexplored in context of computer science.

The key results to note are listed below:

- With regards to research question 4a:
  - As initial cooperation in the population increases, a premium is placed upon lesser tolerance since CC is more likely to be met upon in initial rounds and there is more value in reducing exploitation by initial defectors.
  - As initial defection in the population increases, a premium is placed upon increased tolerance since there is more value in establishing CC cycles from CD/DC plays due to the high likelihood of establishing DD plays initially.
  - As initial cooperation in the population increases, a premium is placed upon moderately responsive emotional characters due to their ability to exploit opponents for one round more than highly responsive emotional characters.
  - As initial defection in the population increases, a premium is placed upon highly responsive emotional characters since they have the highest likelihood of establishing CC plays from CD/DC plays.
- With regards to research question 4b:
  - As the initial percentage of less impressionable agents increases, highly tolerant and highly responsive emotional characters become prevalent due to their high likelihood of establishing CC plays when presented with CD/DC plays.

  - Increasing the initial percentage of highly impressionable agents turns character prevalence into a simple probability game: if there is a higher concentration of emotional characters in a neighbourhood they are likely to gain a foothold and propagate more than emotional characters who are distributed more sporadically throughout the population.

- With regards to research question 4c:

  - Increasing the number of agents in player and comparator sets exaggerates and concentrates emotional character prevalence. Increasing the number of agents in the player set achieves this more so.
  - Highly tolerant emotional characters are admired most and tolerance itself is admired more than responsiveness.

- With regards to research question 5:

  - Irrespective of initial conditions, A3:G1 is the most prevalent emotional character overall. This is due to its high probability of establishing CC cycles regardless of its initial behaviour. A3:G1 is capable of "safely" exploiting opponents i.e. exploitation carries with it no risk of achieving a lower system score as it does with A3:G2.
  - With respect to the above result, total score maximisation is an emergent property of the system since A3:G1 is most proficient at maximisation of the total system score (out of all other emotional characters implemented) and this is the emotional character that always emerges as prevalent (but this ability is not explicitly admired).

As has been mentioned a number of times throughout this chapter, once CC/DD plays are locked into they are never broken out of. This is a point worthy of development and it would be interesting to develop a computational mechanism inspired by emotion that causes periodic defection which does not rely on the arbitrary nature of the "random" and "joss" agents of chapter 6. This would enable the destabilisation of CC plays and force emotional characters to deal with CD/DC plays that emerge. The following question may then be asked: "*what emotional character is most successful when CC gives rise to periodic defection?*". Therefore, the simulations in the next section will look at this in terms of an additional emotion that will be developed: *hope*.

Chapter 8

Hope

Up until this point, the emotion classes of anger, gratitude and admiration have been computationally modelled and implemented to facilitate the investigation of a variety of research questions concerned with how these emotions affect various facets of social interactions in the context of public goods games. In chapter 6, a number of simulations were run which pitted nine emotional characters against each other and against leading strategies from Axelrod's computer tournaments [7]. Results from these simulations show that the highly responsive and highly tolerant emotional character A3:G1 is most proficient at promoting total system scores when playing against Axelrod strategies that periodically defect. Chapter 6 concluded with the finding that, whilst high tolerance and high responsiveness did indeed promote total system scores, they did so at the cost of an individual's score. It was determined that, as long as the individual score of the highly tolerant and highly responsive agent was greater than the individual score that would be achieved through constant DD for a whole game, this reduction in fairness for the good of the system can be considered acceptable.

In chapter 7, I developed the simulations used in chapter 6 to provide experimental evidence illustrating how the characteristics of high tolerance and high responsiveness become proficient in heterogeneous emotional character populations despite various initial conditions being altered.

However, the results in these chapters are limited in that there is no way of asserting that high tolerance and high responsiveness are beneficial in populations where an entire population of agents are endowed with emotions (unlike the simulations in chapter 6) and where periodic defection is possible (unlike the simulations in chapter 7).

The purpose of this chapter is to address the shortcomings of previous chapters by performing a rigorous investigation of how each of the nine basic emotional characters initially defined in chapter 6 handles periodic defection from neighbours in a population of emotional agents whose emotional characters are homogeneous. The periodic defection employed by agents will again be driven by a computational mechanism inspired by studies of human emotion in public goods games and like the previous two chapters, one of the significant aims of this chapter is to identify a viable emotion that may be used to construct this computational mechanism. In this case, *hope* is the emotion class

that will be used as the foundation for this mechanism's implementation which gives the potential for agents to act *greedily*. As with admiration in chapter 7, agents will be endowed with hope in conjunction with the anger and gratitude emotion classes. These emotional characters are distinct from those used in chapter 7 since admiration is not implemented in this set of simulations.

As mentioned, the simulations themselves are composed of homogeneous emotional character populations unlike the heterogeneous emotional character populations used in the simulations run in chapter 7. In this way, the populations implemented can be said to have achieved some form of societal norm with respect to emotional character. Since periodic defection is possible in these simulations, I aim to investigate how successful each of the nine base emotional characters are in the context of the individual and total system scores when the percentage of initial defectors and activation thresholds of hope in the simulations are altered. Furthermore, I also aim to determine if there are any effects upon individual or total system scores caused by altering the activation threshold of hope. This line of enquiry is inspired by the research conducted in chapter 7 (the effects of differing levels of admiration upon character prevalence were analysed) and by Szabó and Thöke's analysis of cooperator densities in populations where the temptation to defect differs [194]. I also investigate whether or not there are any noticeable differences in the behaviour of emotional characters due to differing levels of tolerance and responsiveness when periodic defection is possible.

The structure of this chapter follows that of chapters 6 and 7 with section 8.1 first stating the research questions that this chapter will attempt to answer. Section 8.2 discusses hope by providing justifications for its use in the simulations implemented and how the emotion class will be computationally modelled (eliciting conditions, effects, potential calculation and probability of effect). Moreover, this section will make the link between hope and greed clear since it may seem like a large conceptual leap at the moment to link these two notions together. Section 8.3 describes the emotional characters used in the simulations whilst section 8.4 provides details of the simulations that were run. Section 8.5 first studies the results obtained in the simulations and then goes on to explain how these results emerged. This discussion is then related to the research questions presented in section 8.1 and seeks to provide satisfactory answers to them. Finally, a summary of the chapter is given in section 8.6.

8.1 Research Questions

1. Is hope a suitable emotion to enable agents to periodically defect after CC has been established and why?

2. How should the emotion class of hope be modelled computationally using the emotion model proposed in chapter 5, section 5.3?

3. How should the base emotional characters outlined in section 6.3 of chapter 6 be augmented with hope?

4. What effects, if any, are observed with respect to the individual or total system success of emotional characters when the percentage of initial defectors in a population is varied and why?

5. What effects, if any, are observed with respect to the individual or total system success of emotional characters when periodic defection is less/moderately/more likely and why?

6. When periodic defection is possible, are there any noticeable behavioural features of emotional character populations that are:

   (a) More tolerant than responsive?

   (b) More responsive than tolerant?

   (c) Equally tolerant and responsive?

Note that these research questions will produce results that have not (to the best of my knowledge), been explored by other computer-science researchers that have modelled emotions in public goods games. Therefore, it is impossible to comment upon whether any results are supported/refuted by other pieces of existing research. Essentially, the research performed in this chapter is entirely novel.

8.2 Why Hope?

In both non-human and human populations it is näive to posit that cooperation - after its establishment - will continue indefinitely. In situations where continued CC is a reality, there is an ever present temptation to be greedy when one of the cooperative individuals has more to gain by exploiting a fellow cooperator. This is demonstrated by the negative outcome outlined by the Tragedy of the Commons (see section 2.1.3 of chapter 2) and the preference ordering of outcomes from the perspective of the individual present within the Prisoner's Dilemma game. The temptation to be greedy in the context of Prisoner's Dilemma games is mentioned by a number of researchers. Of particular note are Rapoport and Chammah [154] who, from the perspective of the defector, label the CD play as "temptation" since the defector can receive the maximum individual score in a round by defecting against its cooperative opponent. As mentioned, this link between the term "temptation" and playing as the defector in CD Prisoner's Dilemma games is noted by a sizeable number of researchers including Kuhn and Moresi [107], Szabó and Thöke [194] and Clark and Sefton [31] amongst numerous others.

With particular reference to Ahn et. al., [4], it is proposed that the defector in a CD play can be described as being "greedy" since giving in to the temptation to defect against a cooperative opponent results in an increase of individual pay-off for the defector at the expense of the opponent and the total system. Also of note is Simpson's analysis of gender and cooperation in social dilemmas (particularly the Prisoner's Dilemma) [175],

where it is asserted that greed i.e. the temptation to defect, exists in the Prisoner's Dilemma game and is an independent variable in the experiments outlined.

Since the OCC model [142] is the psychological model of emotion underpinning my own work, greed cannot be considered as a candidate emotion by which a computational mechanism can be translated because it is not considered to be a fully-fledged emotion by Ortony et al. Greed could however be classed as a *characteristic* of an emotion in much the same way as tolerance, responsiveness and impressionability are characteristics of anger, gratitude and admiration respectively. Given that the temptation to defect can be described as being greedy, it is not unreasonable to posit that greed is a characteristic associated with what the OCC model [142] calls a "prospect-based" emotion of which the OCC includes two: hope and fear.

Fear is defined by Ortony et al. as being an emotion that results from displeasure about the prospect of some undesirable event occurring. Hope on the other hand is defined as being an emotion that results from being pleased about the prospect of some desirable event occurring. The event that is the focus of hope and fear is undefined in [142] and may therefore be anything, but in order for the emotion class to be considered rational, the event that elicits the emotion and its effect must be defined and noted in some form of knowledge base. Furthermore, the eliciting conditions and effects of hope or fear must be logically consistent in accordance with the assertions made in section 3.2 of chapter 3. A choice must therefore be made between whether to focus on hope or fear but such a choice is difficult since the OCC model's definitions can be used to justify the use of either in motivating periodic defection. For example: fear may be elicited if a cooperative agent believes that the prospect of its cooperative opponent defecting is likely. In this case, the agent may defect to gain the upper hand. Alternatively, hope may be elicited if a cooperative agent believes that the prospect of defecting against its cooperative opponent will pay off.

I have therefore modelled greed as a characteristic of the OCC's concept of the emotion *hope*. This relationship between hope and greed may seem counter-intuitive to some since hope has positive connotations and greed has negative connotations. However, people can experience hope with respect to negative goals: you can hope for an adversarial work colleague to be removed from his/her position, for example. Certainly in games, one often hopes that the opponent will make a mistake and give one an undeserved win.

The remainder of this section will be devoted to clarifying the relationship between hope and greed and the properties of hope and greed that are important in the context of this thesis. Section 8.2.1 describes the eliciting conditions and effects of hope and section 8.2.2 outlines how hope is computationally modelled for the purposes of the simulations that were run. The aim of these sections is aim to provide clear answers to research questions 1, 2 and 3 posed earlier in section 8.1.

As in chapters 6 and 7, the algorithm for hope implemented in each emotional agent is not detailed specified in detail until section 8.4.4 because additional information regarding the simulation set-up is required to fully understand its operation.

8.2.1 Eliciting Conditions and Effects

A comprehensive analysis of hope as an emotion is undertaken by Lazarus in [109] where it is proposed that hope is elicited by:

> "...a strong desire to be in a different situation than at present, and from the impression that this is possible, either as a result of our own efforts or external forces we do not control."

In the context of hope, the "different situation" mentioned here is defined by Lazarus as being a situation that is more favourable than the one an agent is currently in. Therefore, Lazarus' proposition of hope's eliciting conditions runs parallel to the eliciting conditions proposed by Ortony et al. for hope in [142]. In both cases, the focus of hope is a desirable event or favourable situation that is deemed to be possible from the perspective of the agent. What is interesting about Lazarus' consideration of hope is the discussion of its action-orientation i.e. does the activation of hope in an agent cause that agent to act so as to bring about the desirable event or favourable situation? The overriding answer provided in [109] is that the amount of empirical research regarding this issue is insufficient to make any concrete assertions about this facet of the emotion. Despite this, the assertion that the "different position" (the focus of hope) can result from the efforts of the agent experiencing the emotion, is promising. Further backing for hope's eliciting conditions being focused upon prospective events is provided by Nesse in [133] where it is stated that:

> "Hope and despair exist in the middle realm, when efforts are ongoing but the goal is not yet reached nor recognized as impossible."

Again, hope is elicited in response to a future situation; a prospective event. Unfortunately, as with Lazarus' article [109], the effects of hope are not explored in [133].

Hope's effects upon behaviour are most notably investigated by Snyder in [182] where his "hope theory" is presented. In the context of hope theory, hope is defined as:

> "...the perceived capability to derive pathways to desired goals, and motivate oneself via agency thinking to use those pathways."

Given this definition it is clear that Snyder's consideration of hope is much more effect-orientated than the other works previously discussed. With respect to this definition of hope there are a couple of important concepts to clarify since they have specific definitions in context of Snyder's hope theory. The "pathway" that is derived by hope is determined by means of "pathway thinking", a style of thinking where an agent identifies how to successfully achieve a goal in the future given the current situation. "Agency thinking" is defined by Snyder as "*the perceived capacity to use one's pathways to reach desired goals*" and is the motivational component in hope theory. Essentially, agency thinking identifies specific actions and behaviours that can be utilised to reach

the desired goal. The central argument put forward by hope theory is that hope acts to combine these two styles of thinking (agency and pathway) so that an agent may identify some course of action to achieve a future goal (in the case of hope, the goal is desirable) and then motivate itself to follow this pathway by performing the actions that it has identified as necessary.

What is particularly interesting about Snyder's hope theory is the determination of how much a desirable goal is pursued when it has been identified as being desirable to the agent. In [182] it is proposed that the value of a desirable event or goal is appraised before the event or goal is achieved. In this way, whilst hope motivates both the identification of pathways to achieve desirable events or goals and the implementation of actions to achieve these desirable events or goals, the actions themselves may not be executed. Whether or not these actions are performed is dependent upon the *value* of the goal, as Snyder states:

> "If the imagined outcome of the goal pursuit is sufficiently important to warrant continued mental attention, the person then moves to the event sequence analysis phase."

In other words, if hope is activated towards a desirable event or goal and the value of this event or goal is deemed sufficient to maintain attention then the agent will begin to try to achieve this goal by planning and performing particular actions. It is here where the theory is especially enlightening since Snyder then goes on to differentiate between "high-hope" and "low-hope" individuals. High-hope individuals will essentially have their pathway and agency thinking reinforced, resulting in attention and motivation being sustained with respect to the particular task in hand. Conversely, low-hope individuals do not have their pathway and agency thinking reinforced as much as high-hope individuals and will tend to focus upon other cognitions that are not related to the desirable goal. This reasoning is backed up by a number of laboratory experiments conducted by Snyder and other psychologists including [140], [141], [179], [180] and [181].

8.2.2 Modelling Hope Computationally

Given the research mentioned in section 8.2.1, the task of computationally modelling hope in context of the Prisoner's Dilemma is quite ambiguous and open for interpretation. Since agents are usually assumed to be self-interested I have decided to select an increase in individual score as the desirable state that is hoped for by all agents specifically: agents hope that they may distribute the sucker's pay-off to an opponent and reap the benefits of an increased individual pay-off themselves. Consequently, hope's potential is increased by 1 when an agent and an opponent meet upon CC and hope's potential is reset to 0 should any play other than CC subsequently occur. This eliciting condition was chosen since it is reasonable to propose that, as it becomes more likely that a consistent cycle of CC has been locked into with an opponent, an agent is more

tempted to defect and secure an increase in individual pay-off. Non-CC from an opponent does not increase an agent's hope potential since defecting against an opponent will increase its anger threshold. Under such circumstances, the agent will not benefit from hope's effect since DD is much more likely to be encountered if the agent whose hope is activated periodically defects against an opponent who will defect due to its anger being activated. The idea is to establish goodwill which can be exploited later, rather than to try the patience of one's opponent to breaking point.

Unlike the computational implementations of anger, gratitude and admiration, the activation threshold for all instances of hope in all agents is *constant*. Therefore, hope is always elicited when its potential equals 3 i.e. three consecutive CCs with an opponent must occur for an instance of hope to be elicited with respect to that opponent. The decision to keep hope's activation threshold constant was taken because there is no mention of differing activation thresholds for hope mentioned in any of the research papers considered in section 8.2.1.

The implementation of the hope emotion class in this thesis is the first case where an emotion class' *probability of effect* is varied. Inspiration for this decision came exclusively from Snyder and his consideration of high-hope and low-hope individuals [182]. As explained in section 8.2.1, the difference between high-hope and low-hope individuals is their motivation to both focus on the attainment of a desirable event or goal and as a consequence, their commitment to the actions identified that can achieve this desirable event or goal. Three probabilities of effect (and one control) are modelled in this chapter in terms of hope, augmenting the nine basic emotional characters defined using anger and gratitude (see chapter 6) with three extra sub-classes (reminiscent of how admiration extends the nine basic emotional characters in chapter 7). Details of these new emotional classes are provided in section 8.3 along with a description of the characteristics they create for agents.

8.3 Emotional Characters

Following the clarification of hope's eliciting conditions and effects in a computational context (see section section 8.2.2 above), this section intends to make explicit the emotional characters and associated characteristics that arise from modelling hope in the way proposed.

As outlined in section 8.2.2, there are four probabilities of effect that the emotion may be endowed with: 0, 0.1, 0.5 and 0.9. These probabilities of effect produce four extra emotional characters that each basic emotional character outlined in section 6.3 of chapter 6 may be augmented with: H:0, H:1, H:2 and H:3 [1]. For clarification purposes, the relationships between hope's activation thresholds and these new emotion characters are outlined below:

- H:0 - hope's probability of effect = 0.

[1] **H** is a contraction for ***H****ope*

- H:1 - hope's probability of effect = 0.1.
- H:2 - hope's probability of effect = 0.5.
- H:3 - hope's probability of effect = 0.9.

The relationships above are interpreted thus: when hope is elicited, its effect (to periodically defect) is manifest 10% of the time in those agents whose emotional characters are extended with H:1; 50% of the time for emotional characters extended with H:2; 90% of the time for emotional characters extended with H:3; and 0% of the time for emotional characters extended with H:0. Altering these probabilities in the population is the focus of research question 5 (see section 8.1) and their effects upon individual and total system scores are discussed in section 8.5.

Like admiration in chapter 7, hope is intended to extend the nine basic emotional characters defined in context of anger and gratitude. This means that the total number of emotional characters possible to be implemented is extended to a maximum of thirty-six. So, if the two-dimensional emotional character matrix from section 6.3 of chapter 6 (see table 6.1) is again used as a basis, a third dimension may be added that extends each basic emotional character to include the four extra emotional characters noted above. For example: if A1:G1 is considered then, with the addition of hope, emotional characters A1:G1:H0, A1:G1:H1, A1:G1:H2 and A1:G1:H3 are possible. This of course produces additional emotional characteristics (see section 5.4.1 of chapter 5) in addition to tolerance and gratitude. Since hope motivates agents to defect against cooperative opponents, its associated characteristic in this case is *greed*. Also, because hope is modelled with having four probabilities of effect, an agent may be therefore characterised as being either highly greedy, moderately greedy, less greedy or non greedy. Hope's probabilities of effect and the associated characteristics are detailed below for reference using A1:G1 as an example:

- A1:G1:H0 - Less tolerant, highly responsive and non greedy.
- A1:G1:H1 - Less tolerant, highly responsive and less greedy.
- A1:G1:H2 - Less tolerant, highly responsive and moderately greedy.
- A1:G1:H3 - Less tolerant, highly responsive and highly greedy.

8.4 Simulation Details

Again, the simulation set-up used in this chapter is largely similar to that described in section 6.4 of chapter 6 i.e. an iterated Prisoner's Dilemma game is played in the context of a MAS simulation. The pay-off matrix illustrated in section 2.1.2 of chapter 2 (see table 2.1) is still used and the simulation environment is still based upon a two-dimensional grid.

The main difference in this simulation regards the removal of the player and comparator sets used in the investigation of admiration in chapter 7. This decision was made since I am not interested in the effects of altering these neighbourhoods with respect to hope and so there is no requirement for them in the simulations run in this chapter. Consequently, each agent plays against the two agents directly adjacent to it; doubling the number of interactions for the size of simulation without over complicating the analysis.

This section is segregated into four parts: important details regarding the initial set-up of simulations are contained in section 8.4.1; particular details of agents i.e. initial behaviour determination, types of agents implemented and other actions performed will be discussed in section 8.4.2; section 8.4.3 outlines the progression of the simulation both on a round-to-round and game-to-game basis; section 8.4.4 delineates the hope algorithm implemented in every emotional agent in the simulations.

8.4.1 Simulation Set-up

The simulation environment used in this chapter is nearly identical to that used previously in chapter 7. The edges of the environment still "wrap" so that all agents play against an equal number of opponents. As with the simulation environment used in chapter 7, 338 agents make up the population and all agents are endowed with emotional characters. The decision to maintain a population of agents rather than pitting two agents of the same emotional character together was made so that answers could be provided for research questions 4-6 (see section 8.1). Providing answers to these questions necessitates interactions between more than two agents and since 338 agents provided adequate results in the simulations run in chapter 7, I simply decided to maintain this number for the simulations used in this chapter.

8.4.2 Agent Details

Like the emotional agents implemented in chapters 6 and 7, the emotional agents in this simulation utilise an initial behaviour setting to generate their behaviour until either anger, gratitude or hope is activated. After any of these emotions are activated, the agent's potentials for anger, gratitude and hope will never all be reset to 0 before the game ends. Therefore, after activation of one of these emotions, an agent's behaviour is continuously driven by its emotional state meaning that the initial behaviour setting is redundant and no longer needed. As also mentioned in section 8.4.1, the percentage of initial cooperators and defectors in the population can be set although it is not possible to specify such values on an individual agent level. Once the desired percentage of initial cooperators and defectors in the population has been selected, each agent is randomly assigned an initial behaviour so as to prevent any bias occurring with respect to individual or total score success due to intentional patterns of initial behaviour distribution.

8.4.3 Simulation Progression

A typical round in the simulation proceeds as described below.

1. Agents in the population are generated with an emotional character and are randomly assigned an initial behaviour. The percentage of initial cooperators and defectors in the population equals the percentages specified in the simulation's set-up files.

2. Agents consult their current emotional state if an emotion is active to determine their behaviour in this round or, if no emotion is active, their initial behaviour setting.

3. Agent's emotional state is updated by reacting to the opponent's behaviour in this round and pay-offs are distributed by each agent as per the pay-off distribution for the Prisoner's Dilemma (see table 2.1 located in section 2.1.2 of chapter 2).

4. Round ends, return to step 3 and repeat until 500 rounds have been played.

In order to answer the research questions detailed in section 8.1 a number of scenarios were designed, implemented and run. Five general scenarios were run for each of the thirty-six emotional character types (a summary of these scenarios can be found in table 8.1) and each scenario was repeated five times for each emotional character population. In total, data for 900 simulations was collected and I assert that such a volume of information is adequate to allow for a robust, valid discussion and to allow the proposal of conclusions with a high degree of confidence. Scenario 6 acts as a control scenario since DD is never destabilised unlike CC which can be destabilised when greed is present within the population.

TABLE 8.1: Scenario Number and Initial Percentage of Cooperators/Defectors in Population.

Scenario #	**% of Initial Cooperators in Population**	**% of Initial Defectors in Population**
1	100	0
2	80	20
3	60	40
4	40	60
5	20	80
6	0	100

8.4.4 Hope Algorithm

1. After each round, agent x analyses its opponent, y's, behaviour in the round.

(a) If y defected, x's hope potential towards y is reset to 0 (x's hope instance for y is deactivated if currently active).

(b) If y cooperated:

i. If x's hope instance for y is not currently active, x checks own behaviour in the round just played:

A. If x cooperated, increase hope potential towards y by 1.

B. If x defected, reset hope potential towards y to 0.

ii. If x's hope instance for y is currently active, x does nothing (x's hope instance for y remains active).

2. x checks hope potential for y. If x's hope potential for y equals x's hope activation threshold, x's hope instance for y is activated (if currently not active) and the following occurs:

(a) x randomly selects an integer between 1 and 10.

i. If integer is less than or equal to x's probability of hope's effect being manifest, x will defect against y in the next round.

ii. If integer is greater than or equal to x's probability of hope's effect being manifest, y will cooperate with y in the next round.

8.5 Results and Analysis

In this section I detail all results obtained from running the simulations outlined in section 8.4.1. Prior to commencing any detailed analysis, it is necessary for me to first outline how the four possible play probabilities between an agent and its opponent CC, CD, DC and DD can be altered by varying the scenario number and the level of greed present within the population. Accordingly, the results obtained in these simulations are similar to those obtained in chapter 7 since these probabilities have a significant effect upon the plays that develop between agents and their opponents in populations. Therefore, outlining such probabilities is integral to understanding the conclusions drawn from an analysis of the results. Table 8.2 illustrates how *initial* play probabilities are affected by changing the simulation scenario.

Table 8.2: The effect of scenario upon initial play probability.

Scenario #	**Prob. of CC**	**Prob. of CD**	**Prob. of DC**	**Prob. of DD**
1	1	0	0	0
2	0.64	0.16	0.16	0.04
3	0.36	0.24	0.24	0.16
4	0.16	0.24	0.24	0.36
5	0.04	0.16	0.16	0.64
6	0	0	0	1

Since each game contains a homogeneous emotional character population (for reasons outlined in the introduction of this chapter, it is possible to determine the precise effect exerted on the probabilities of the four plays outlined above occurring *after* hope has been activated by varying the probability of hope's effect being manifest in an agent. These effects are outlined in table 8.3. It should be noted that the effects of having non-greedy emotional character populations are not detailed since, in this case, only the probabilities outlined in table 8.2 are relevant.

TABLE 8.3: The effect of hope's likelihood of activation upon the probability of CC, CD, DC and DD occurring between an agent and an opponent.

Character	**Prob. of CC**	**Prob. of CD**	**Prob. of DC**	**Prob. of DD**
H:1	0.81	0.09	0.09	0.01
H:2	0.25	0.25	0.25	0.25
H:3	0.01	0.09	0.09	0.81

It is important to again highlight that scenario 6 acts as a control for these simulations i.e. the scores achieved by all agents within each emotional character population are equal (1000: 500 from each opponent). As such, results from this scenario are not discussed in detail in this section since altering the probability that hope's effect is manifest when activated exerts no effect upon any emotional character population.

Note also that the activation thresholds of anger and gratitude i.e. the ratio of responsiveness to tolerance, embodied by an emotional character is a crucial determinant of success in these simulations and is the linchpin that holds the results observed and the explanations provided together. Emotional characters A2:G1, A3:G1 and A3:G2, whose activation threshold for gratitude is less than their activation threshold for anger, always convert CD/DC plays into CC (as seen in table 8.4) earning agents 6 points in each round.

Emotional characters A1:G1, A2:G2 or A3:G3, whose activation thresholds for gratitude and anger are equal, maintain CD/DC plays for the rest of a game after their establishment (see table 8.5) earning agents 5 points in irregular intervals.

Emotional characters A1:G2, A1:G3 or A2:G3, whose activation threshold for gratitude is greater than its activation threshold for anger, always convert CD/DC plays into DD (see table 8.6) earning agents 2 points in each round.

This section is broadly split into two subsections: section 8.5.1 identifies the effects of altering the percentage of initial defectors upon emotional character success. This consideration entails an analysis of the effect of greed upon emotional character success since *all* emotional character populations are considered. Therefore, section 8.5.1 aims to provide an answer to research questions 4 and 5. Section 8.5.2 comments upon the effects of tolerance and responsiveness with respect to the success (both individual and group success) of emotional characters in the context of different tolerance and responsiveness ratios when agents are capable of greed.

TABLE 8.4: Play histories illustrating how increasing the activation threshold for anger and reducing the activation threshold for gratitude results in emotional characters A2:G1, A3:G1 and A3:G2 establishing CC plays from CD/DC plays.

	A2:G1		**A3:G1**		**A3:G2**	
Round #	*Agent x*	*Agent y*	*Agent x*	*Agent y*	*Agent x*	*Agent y*
1	C	D	C	D	C	D
2	C	C	C	C	C	D
3	C	C	C	C	C	C
Individual Score	*6*	*11*	*6*	*11*	*3*	*13*

TABLE 8.5: Play histories illustrating how keeping the activation thresholds for anger and gratitude equal results in emotional characters A1:G1, A2:G2 and A3:G3 maintaining CD/DC plays after their establishment.

	A1:G1		**A2:G2**		**A3:G3**	
Round #	*Agent x*	*Agent y*	*Agent x*	*Agent y*	*Agent x*	*Agent y*
1	C	D	C	D	C	D
2	D	C	C	D	C	D
3	C	D	D	C	C	D
4	D	C	D	C	D	C
5	C	D	C	D	D	C
6	D	C	C	D	D	C
Individual Score	*15*	*15*	*10*	*20*	*15*	*15*

TABLE 8.6: Play histories illustrating how decreasing the activation threshold for anger and increasing the activation threshold for gratitude results in emotional characters A1:G2, A1:G3 and A2:G3 establishing DD plays from CD/DC plays.

	A1:G2		**A1:G3**		**A2:G3**	
Round #	*Agent x*	*Agent y*	*Agent x*	*Agent y*	*Agent x*	*Agent y*
1	C	D	C	D	C	D
2	D	D	D	D	C	D
3	D	D	D	D	D	D
4	D	D	D	D	D	D
5	D	D	D	D	D	D
6	D	D	D	D	D	D
Individual Score	*5*	*10*	*5*	*10*	*4*	*14*

8.5.1 Effects of Initial Defection Percentages and Greed Upon Emotional Character Success

To ascertain the effects of initial defection percentages and likelihood of greed within emotional character populations three variables of interest common to all emotional character populations were identified. These variables are each considered individually

so that answers to research questions 4 and 5 (see section 8.1) can be provided. These variables along with their descriptions and how each was obtained, are presented below.

- *Maximum individual score*: The maximum individual score for an emotional character population is derived by selecting the maximum individual score within each emotional character population after the final round of each game has been played. With respect to the five repeats of each scenario, the highest individual score obtained by each emotional character population is identified and recorded. After these scores have been collected for each emotional character population, they are ordered to determine how successful each emotional character population is at maximising the score of its individuals in context of each scenario. To clarify: if the majority of maximum individual scores over the five repeats of a scenario belong to emotional character n then it may be posited that, given the conditions outlined by that scenario, emotional character n agents achieve the highest individual scores possible compared with all other emotional character populations.

- *Minimum individual score*: See maximum individual score derivation procedure above but replace "maximum" with "minimum".

- *Total system score*: The total system score for an emotional character population is derived by summing together the scores of each individual within each character population after the final round of each game has been played. These scores are then ordered in the same way that maximum and minimum individual scores are.

These criteria were chosen since ascertaining the maximum/minimum individual scores for a population allows for the absolute determination of how well a particular emotional character promotes individual scores and guards against individual score reduction. Standard deviation is not used since this metric would lend itself better to a determination of fairness rather than a commentary such as the one desired. Analysing the maximum total score of each character population allows me to comment upon the success of each emotional character population at a macro level.

These criteria also permit the collection of information pertaining to how successful an emotional character is with respect to establishing and maintaining cooperation. By analysing against these criteria I aim to identify and comment upon specific features of emotional characters that cause the results obtained to be achieved which will help to provide answers for research question 6.

This section is therefore divided into three subsections: 8.5.1.1, 8.5.1.2 and 8.5.1.3. These sections respectively consider maximum individual scores, minimum individual scores and total system scores when there are differing percentages of initial defectors and greed levels within the population. Some may wonder why a discussion of fairness is not included in this section since maximum and minimum individual scores would seem to provide a lead in to such a discussion. However, this discussion is postponed until section 8.5.2.3 where the behavioural features of emotional characters with different

tolerance and responsiveness ratios is dealt with. In section 8.5.2.3 I can succinctly analyse fairness in context of the three tolerance and responsiveness groups described earlier rather than considering each emotional character in turn.

In each repeat for each scenario the maximum individual score, minimum individual score and maximum total system score for each emotional character population in each scenario was identified. These scores were then ranked along with the emotional characters that achieved them and placings were determined. So, if an agent in the A3:G1 population achieved the highest individual score in any of the five repeats of each scenario when compared to all other emotional character populations in the same scenario then that emotional character places first.

The process of ordering particular scores (described by the bullet points above) enables "scores" or "points" to be distributed to each emotional character. The use of a consistent scoring mechanism such as this enables the success of each emotional character population in context of each of the three criteria identified to be determined and compared. Such a feature is important since different greed likelihoods exist and I require a way to be able to compare populations of different greed likelihoods against each other (so that research questions 4 and 5 may be answered).

Points are allocated to each emotional character population out of a total of 36 (since there are this many emotional characters simulated in context of each scenario). Ergo, if the emotional character n population contains an agent that has achieved the highest maximum individual score in one of the repeats of the scenario being considered then, emotional character n is awarded 36 points. The emotional character population that contains the next highest scoring agent places second is awarded 35 points and so on. If there is a draw in a particular place then the emotional character populations that have drawn are awarded an equal number of points. for example: if emotional character populations n and m both contain an agent who has achieved an equal maximum individual score (entailing that n and m both place first) then populations n and m are awarded 36 points each.

This results in the production of five tables for each variable of interest (maximum individual score, minimum individual score and maximum total system score) for each of the five scenarios implemented. These tables enable me to comment upon the effect of initial defection rates upon emotional character success. Only five tables are produced due to scenario 6's status as a control group: all individual scores are equal in this scenario and commenting upon such scores yields no interesting points. Within the tables constructed, emotional characters are further grouped into their respective likelihoods of greed so that comment can be passed regarding the effect of greed upon the success of emotional characters in context of the three criteria outlined.

8.5.1.1 Maximum Individual Score

Results The lowest maximum individual score obtainable by any agent is 3000, achieved by an agent that establishes and maintains CC from the very first round with both its

opponents for 500 rounds. In theory, the highest maximum score that an individual agent can achieve in one game is 5000, achieved by an agent distributing the sucker's pay-off to both of its opponents in every round for the entire game. However, such a score is not achievable in the simulation because, depending upon how tolerant the agent is, the sucker's pay-off will only be accepted by an agent up to three times.

Tables 8.7, 8.8, 8.9, 8.10 and 8.11 show the placing of emotional characters (organised by greed likelihood) with respect to scenarios 1-5. Note that in table 8.7, there is no column for emotional characters whose likelihood of greed is non-existent. This is because all individuals in these emotional character populations score 3000 since all individuals cooperate with their opponents initially. This ultimately means that hope is never triggered, so periodic defection never has a chance to occur; in the context of this scenario and for emotional characters who will never act greedily, CC plays never break once established.

TABLE 8.7: Emotional character placing based upon maximum individual scores obtained in scenario 1 organised by the probability of hope's effect being manifest after activation.

Position	**H:1**	**H:2**	**H:3**
1	A3:G1	A3:G3	A3:G1
2	A2:G1	A2:G2	A3:G3
3	A3:G3	A1:G1	A2:G2
4	A2:G2	A3:G1	A1:G1
5	A1:G1	A2:G1	A3:G2
6	A3:G2	A3:G2	A2:G1
7	A2:G3	A2:G3	A2:G3
8	A1:G3	A1:G2, A1:G3	A1:G2, A1:G3
9	A1:G2		

TABLE 8.8: Emotional character placing based upon maximum individual scores obtained in scenario 2 organised by the probability of hope's effect being manifest after activation.

Position	**H:0**	**H:1**	**H:2**	**H:3**
1	A3:G2	A3:G1	A3:G3	A3:G3
2	A2:G1, A3:G1	A2:G1	A2:G2	A1:G1, A2:G2
3	A1:G1, A1:G2, A1:G3, A2:G2, A2:G3, A3:G3	A3:G3	A1:G1	A3:G2
4		A2:G2	A3:G1	A3:G1
5		A1:G1	A2:G1	A2:G1
6		A3:G2	A3:G2	A2:G3
7		A2:G3	A2:G3	A1:G2, A1:G3
8		A1:G2	A1:G3	
9		A1:G3	A1:G2	

TABLE 8.9: Emotional character placing based upon maximum individual scores obtained in scenario 3 organised by the probability of hope's effect being manifest after activation.

Position	**H:0**	**H:1**	**H:2**	**H:3**
1	A3:G2	A3:G1	A3:G3	A3:G3
2	A2:G1, A3:G1	A2:G1	A2:G2	A1:G1, A1:G2
3	A1:G1, A1:G2, A1:G3, A2:G2, A2:G3, A3:G3	A1:G1, A3:G3	A1:G1	A3:G1
4		A2:G2	A3:G1	A3:G2
5		A3:G2	A2:G1	A2:G1
6		A1:G2	A3:G2	A2:G3
7		A2:G3	A2:G3	A1:G3
8		A1:G3	A1:G3	A1:G2
9			A1:G2	

TABLE 8.10: Emotional character placing based upon maximum individual scores obtained in scenario 4 organised by the probability of hope's effect being manifest after activation.

Position	**H:0**	**H:1**	**H:2**	**H:3**
1	A3:G2	A3:G1	A3:G3	A3:G3
2	A2:G1, A3:G1	A2:G1	A2:G2	A1:G1, A2:G2
3	A1:G1, A1:G2, A1:G3, A2:G2, A2:G3, A3:G3	A3:G3	A1:G1	A2:G1
4		A1:G1	A3:G1	A3:G1
5		A2:G2	A2:G1	A3:G2
6		A3:G2	A3:G2	A2:G3
7		A2:G3	A2:G3	A1:G2, A1:G3
8		A1:G3	A1:G3	
9		A1:G2	A1:G2	

No Greed The placing of non-greedy emotional characters is consistently determined according to their tolerance/responsiveness ratios across scenarios 2-5. As identified in chapter 7, the success of A3:G2 with respect to maximisation of individual score is due to its ability to exploit opponents for two rounds before locking into CC plays with them. A2:G1 and A3:G1 on the other hand establish CC after 1 round of CD/DC play. Therefore, if the defecting agent in an initial CD/DC play is A3:G2, this agent is capable of maximising its own individual score by administering the sucker's pay-off *twice* to its opponent before establishing CC. If the defecting agent in a CD/DC play is A2:G1 or A3:G1, this agent may only administer the sucker's pay-off once. To make this advantage clearer, consider the example outlined previously in table 8.4.

All other emotional characters are not capable of exploiting an opponent before establishing CC (see tables 8.5 and 8.6) therefore, the highest score achieved by these

TABLE 8.11: Emotional character placing based upon maximum individual scores obtained in scenario 5 organised by the probability of hope's effect being manifest after activation.

Position	**H:0**	**H:1**	**H:2**	**H:3**
Position	*None*	*Low*	*Mod.*	*High*
1	A3:G2	A3:G1	A3:G3	A3:G3
2	A2:G1, A3:G1	A2:G1	A1:G1, A2:G2	A1:G1, A2:G2
3	A1:G1, A1:G2, A1:G3, A2:G2, A2:G3, A3:G3	A2:G2	A3:G1	A3:G2
4		A1:G1	A2:G1	A3:G1
5		A3:G3	A3:G2	A2:G1
6		A3:G2	A2:G3	A2:G3
7		A2:G3	A1:G3	A1:G2
8		A1:G3	A1:G2	A1:G3
9		A1:G2		

emotional character populations is 3000. Such a score is achieved when two agents cooperate initially and this CC is maintained for the remainder of the game. Despite initial cooperator and defector percentages, CC is *always* possible (no matter how unlikely) and since each scenario is repeated 5 times, there is always an agent pair in the population who cooperate with each other initially.

Less Greed The placing of less-greedy emotional characters with respect to their tolerance and responsiveness ratios is mostly consistent across scenarios 1-5. From this it can be determined that it is the tolerance and responsiveness ratios of emotional characters that determine the individual success of agents, but why do these ratios enable the patterns of success shown?

As explained previously and by table 8.4, A2:G1 and A3:G1 place in the top two positions since they are able to quickly establish CC from CD/DC plays (they are the quickest emotional characters to do so out of all the emotional characters implemented). Since the chances of two opponents mutually defecting against one another is so low (0.01, see table 8.3) after hope has been activated for less greedy agents, there is less of a risk associated with re-establishing CC after destabilisation (CC for three rounds triggers hope giving a chance for agents to establish DD). Therefore, an A2:G1 or A3:G1 agent that is the defecting agent in a CD/DC play is able to distribute the sucker's pay-off, re-establish CC in the next round and activate hope 3 rounds later without incurring any serious risk of establishing a DD play. The higher placement of A3:G1 over A2:G1 is due to A3:G1's increased tolerance that enables the re-establishment of CC and avoidance of a retaliatory sucker's pay-off after exploiting an opponent for a maximum of two rounds. A2:G1 agents however are only capable of exploiting opponents for a maximum of one round if CC is to be re-established without the defector incurring a sucker's pay-off.

The placings of A1:G1, A2:G2 and A3:G3 is caused by their inability to establish CC following CD/DC plays. As mentioned in the discussion of non-greedy emotional character placings above, the characteristic behaviour of the agents following CD/DC play is to maintain these plays after their establishment (see table 8.5). This is a safe option when there is a chance of DD being established after hope has been activated. However, because the chance of DD occurring is so low, there is no real premium upon such a characteristic hence these emotional characters do not place higher.

A3:G2's placing is very interesting since this character prevails when greed has no probability of occurring in a population but it does not achieve such success when greed is present. When compared to A2:G1 and A3:G1, the reduced responsiveness of A3:G2 appears to encumber agents with respect to maximisation of individual scores. If two agents, x and y, are A2:G1 or A3:G1 and x suckers y due to greed in round n, and y retaliates by suckering x due to anger in round $n + 1$, then x's cooperation in round $n + 1$ (due to the periodic defection produced by greed) will re-establish CC between x and y in round $n + 2$. If x and y are A3:G2 then y would require two cooperations from x resulting in the A3:G2 agent x achieving a lower individual score than the A2:G1 or A3:G1 agent x. Table 8.12 illustrates the situation described to make the explanation more understandable and transparent (plays in bold are performed when greed is activated for that agent).

TABLE 8.12: How the increased responsiveness of A3:G1 enables quicker establishment of CC plays following retaliatory exploitation when compared to the reduced responsiveness of A3:G2.

	A3:G1		**A3:G2**	
Round #	*Agent x*	*Agent y*	*Agent x*	*Agent y*
n	**D**	C	**D**	C
$n + 1$	**C**	D	**C**	D
$n + 2$	C	C	C	D
$n + 3$	C	C	C	C
Individual Score	*11*	*11*	*8*	*13*

A3:G2's reduced success compared to A1:G1, A2:G2 and A3:G3 is also quite interesting. It would appear that, when there is a chance of periodic defection occurring, reduced responsiveness (compared to increased tolerance) is not a safe option because this can potentially result in DD being established between agents, unless it is counterbalanced with equal tolerance. Indeed, maintaining CD/DC plays is much more lucrative from the standpoint of individual scores than risking the establishment of DD for relatively few exploitations (even when periodic defection is unlikely). Furthermore, the certainty of A1:G2, A1:G3 and A2:G3 establishing DD from CD/DC plays (shown in demonstrated in table 8.6) is also their downfall.

Moderate Greed Moderately greedy emotional characters have a 0.5 chance of defecting after hope is activated (see table 8.3) therefore, since all four play scenarios are equally likely to occur. Consequently, there is an increased premium on emotional characters that can avoid establishing CC/DD. The only characters that are guaranteed to do this are those whose tolerance/responsiveness levels are equal i.e. A1:G1, A2:G2 and A3:G3 (see table 8.5).

Consequently, it is observed in tables 8.7, 8.8, 8.9, 8.10 and 8.11 that emotional characters A1:G1, A2:G2 and A3:G3 consistently place in the top 3 positions. A2:G1, A3:G1 and A3:G2 now consistently place in positions 4-6 since CC promotes maximum individual scores more than DD. However, CC now entails a credible risk of DD plays being established after activation of hope.

A1:G2, A1:G3 and A2:G3 are consistently the least successful in all scenarios. The actual placing of these characters is of no consequence because they appear to shift positioning in their groups from scenario to scenario whilst the actual character groupings (in terms of tolerance and responsiveness ratios) remain consistent. The only explanation that can be offered for this is pure chance: some agents may encounter more favourable game conditions than others.

High Greed With regards to scenario 1 (see table 8.7), the highly greedy A3:G1 agent places first whilst A1:G1, A2:G2 and A3:G3 place second, third and fourth respectively. A3:G1's success is due in part to its high tolerance however, it does appear as though chance plays a large part in this character's success. Since the actual outcomes are determined by a binomial probability distribution, some agents may encounter more favourable conditions that others. Regardless of this, by looking at a portion of the actual play history of the highest scoring agent in A3:G1's population (see table 8.13) it can be seen how advantageous it is to be highly tolerant and highly responsive (plays in bold are performed when greed is activated for that agent).

As can be seen in table 8.13, the initial 3 rounds of cooperation trigger hope in both agents, x and y. In round 4 however, y manages to cooperate even though its chance of doing so is 0.1 (see table 8.3). If this did not occur then it would be likely that the same situation occurs as in round 12: both agents defect and deactivate each other's hope (x defects in round 13 due to y's defection in rounds 7, 8 and 12). Should this occur consistently then a play history would be obtained where, for the first 12 rounds, 9 would consist of CC and 3 would consist of DD. This final DD would then go on to establish and maintain DD for the remainder of the game. A1:G1, A2:G2 and A3:G3 follow a similar pattern i.e. if on round 4 one agent manages to cooperate then a CD/DC play is established. Otherwise, these characters will establish defection plays after 4 (A1:G1), 8 (A2:G2) or 12 (A3:G3) rounds.

Character placings for scenarios 2-5 are identical (see tables 8.8, 8.9, 8.10 and 8.11): A1:G1, A2:G2 and A3:G3 place highest; A2:G1, A3:G1 and A3:G2 place immediately

Table 8.13: How the increased tolerance of A3:G1 facilitates the maximisation of an individual's score even in highly-greedy populations.

	A3:G1	
Round #	*Agent x*	*Agent y*
1	C	C
2	C	C
3	C	C
4	**D**	**C**
5	**D**	C
6	**D**	C
7	**D**	D
8	C	D
9	C	C
10	C	C
11	C	C
12	**D**	**D**
13	D	C
14	C	C

below; A1:G2, A1:G3 and A2:G3 place lowest (with A2:G3 placing the highest out of this subset of emotional characters).

Equally tolerant and responsive emotional characters place highest since there is now an increased premium on maintaining CD/DC plays and avoiding the establishment of CC (as explained previously in the context of moderate greed). A3:G3 places the highest out of the equally tolerant and responsive emotional characters since an A3:G3 agent that defects in a CD/DC play earns 15 points over 3 rounds whereas a defecting A1:G1 or A2:G2 agent only earns 10 points over 3 rounds in the same situation.

A2:G3 places highest out of the more tolerant than responsive emotional characters since an initially cooperative agent with this emotional character requires two defections before its disposition switches from cooperation to defection. This increased tolerance (relative to A1:G2 and A1:G3) postpones the establishment of DD for one round more than A1:G2 and A1:G3 and increases the individual pay-off for the defector agent in a CD/DC play.

8.5.1.2 Minimum Individual Score

Results The lowest minimum score that can be theoretically achieved is 0; obtained when an agent receives the sucker's pay-off for an entire game. However, as explained in the introduction to section 8.5.1.1, no emotional character will tolerate receiving the sucker's pay-off more than 3 times before establishing a DD play or the play of the agent and its opponent is switched (as in TFT). Therefore, the lowest score possible is 994, obtained by receiving the sucker's pay-off for 3 rounds from both opponents followed by the establishment of DD with both opponents and maintenance of this play for the remainder of the game.

Tables 8.14, 8.15, 8.16, 8.17 and 8.18 show the placing of emotional characters (organised by greed likelihood) with respect to scenarios 1-5. To ensure that the tables are interpreted correctly some explanation is required: emotional characters that place in first position in these tables have obtained the most positive minimum individual scores, emotional characters that place in last position have obtained the most negative individual scores.

As with table 8.7 in section 8.5.1.1, table 8.14 contains no column for emotional characters whose likelihood of greed is non-existent. This is because all individuals in these emotional character populations score 3000 as a result of all individuals cooperating with their opponents initially. This ultimately means that hope is never triggered, so periodic defection never has a chance to occur; in the context of this scenario and for emotional characters who will never act greedily, cooperation plays never break once established.

TABLE 8.14: Emotional character placing based upon minimum individual scores obtained in scenario 1 organised by the probability of hope's effect being manifest after activation.

Position	**H:1**	**H:2**	**H:3**
1	A3:G1	A3:G2	A3:G2
2	A3:G2	A3:G1	A3:G1
3	A2:G1	A2:G1, A3:G3	A2:G1, A3:G3
4	A3:G3	A2:G2, A2:G3	A2:G2, A2:G3
5	A2:G2	A1:G1, A1:G2, A1:G3	A1:G1, A1:G2, A1:G3
6	A2:G3		
7	A1:G2		
8	A1:G3		
9	A1:G1		

TABLE 8.15: Emotional character placing based upon minimum individual scores obtained in scenario 2 organised by the probability of hope's effect being manifest after activation.

Position	**H:0**	**H:1**	**H:2**	**H:3**
1	A1:G1, A2:G1, A2:G2, A3:G1, A3:G2, A3:G3	A1:G1, A2:G1, A2:G2, A3:G1, A3:G2, A3:G3	A1:G1, A2:G1, A2:G2, A3:G1, A3:G2, A3:G3	A1:G1, A2:G1, A2:G2, A3:G1, A3:G2, A3:G3
2	A1:G2, A1:G3	A1:G2, A1:G3	A1:G2, A1:G3	A1:G2, A1:G3
3	A2:G3	A2:G3	A2:G3	A2:G3

No Greed For scenarios 2-5, non-greedy emotional character placement is consistent (see tables 8.15, 8.16, 8.17 and 8.18) with the more responsive than tolerant group of emotional characters i.e. A1:G2, A1:G3 and A2:G3, performing worst.

A2:G3 scores the lowest of these characters since individuals in this population can receive up to 2 sucker's pay-offs before establishing DD and maintaining them for the

Table 8.16: Emotional character placing based upon minimum individual scores obtained in scenario 3 organised by the probability of hope's effect being manifest after activation.

Position	**H:0**	**H:1**	**H:2**	**H:3**
1	A1:G1, A2:G1, A2:G2, A3:G1, A3:G2, A3:G3	A1:G1, A2:G1, A2:G2, A3:G1, A3:G2, A3:G3	A1:G1, A2:G1, A2:G2, A3:G1, A3:G2, A3:G3	A1:G1, A2:G1, A2:G2, A3:G1, A3:G2, A3:G3
2	A1:G2, A1:G3	A1:G2, A1:G3	A1:G2, A1:G3	A1:G2, A1:G3
3	A2:G3	A2:G3	A2:G3	A2:G3

Table 8.17: Emotional character placing based upon minimum individual scores obtained in scenario 4 organised by the probability of hope's effect being manifest after activation.

Position	**H:0**	**H:1**	**H:2**	**H:3**
1	A1:G1, A2:G1, A2:G2, A3:G1, A3:G2, A3:G3	A1:G1, A2:G1, A2:G2, A3:G1, A3:G2, A3:G3	A1:G1, A2:G1, A2:G2, A3:G1, A3:G2, A3:G3	A1:G1, A2:G1, A2:G2, A3:G1, A3:G2, A3:G3
2	A1:G2, A1:G3	A1:G2, A1:G3	A1:G2, A1:G3	A1:G2, A1:G3
3	A2:G3	A2:G3	A2:G3	A2:G3

Table 8.18: Emotional character placing based upon minimum individual scores obtained in scenario 5 organised by the probability of hope's effect being manifest after activation.

Position	**H:0**	**H:1**	**H:2**	**H:3**
1	A1:G1, A2:G1, A2:G2, A3:G1, A3:G2, A3:G3	A1:G1, A2:G1, A2:G2, A3:G1, A3:G2, A3:G3	A1:G1, A2:G1, A2:G2, A3:G1, A3:G2, A3:G3	A1:G1, A2:G1, A2:G2, A3:G1, A3:G2, A3:G3
2	A1:G2, A1:G3	A1:G2, A1:G3	A1:G2, A1:G3	A1:G2, A1:G3
3	A2:G3	A2:G3	A2:G3	A2:G3

rest of the game. Indeed, whilst this is both the cause of A2:G3's success with respect to maximum individual score in context of the more tolerant than responsive emotional characters, it is also its downfall. Individuals in populations of A1:G2 and A1:G3 agents only receive the sucker's pay-off once before establishing and maintaining DD for the rest of the game.

The remaining emotional characters contain individuals in their populations that score a minimum of 1000, achieved by establishing and maintaining DD with both opponents from the initial round until the end of the game. As explained in the discussion of non-greedy emotional characters in section 8.5.1.1: A1:G1, A2:G1, A2:G2, A3:G1, A3:G2, A3:G3 are incapable of establishing DD from CD/DC plays so no free-riding can occur before DD plays are established. No other agents in these emotional character populations achieve a worse score than 1000 since it is not possible for any agents in these populations to establish DD with opponents after receiving the sucker's pay-off.

Less Greed With reference to table 8.14, the first and only occurrence of distinct emotional character grouping occurs in scenario 1 with regards to minimum individual scores. These emotional character groupings would appear to be based upon tolerance and responsiveness ratios with the exception of A1:G1's placing in this scenario. The individual score responsible for this low placing is 1012 and its acquisition is due to a combination of intolerance and sheer bad luck. Analysing the play history of the agent who achieved this score, x, reveals that x mutually cooperates for 3 rounds with both its opponents, y and z, then all players i.e. x, y and z, all defect when their hope is activated. This is extremely unlikely to occur but is not impossible. Due to the intolerance of A1:G1, this DD between both opponents causes DD plays to be locked into for the remainder of the game.

A1:G2, A1:G3 and A2:G3 place above A1:G1 in scenario 1 since these emotional characters lock into defection plays later than A1:G1 does in the play history outlined above (it would appear that no individuals in these character populations are as unlucky as A1:G1). A3:G3 and A2:G2 place in positions 4 and 5 respectively due to their TFT play strategies that ensure the maintenance of CD/DC plays (and thus the avoidance of DD) if they are established when greed occurs. A2:G1, A3:G1 and A3:G2 place in 1st, 2nd and 3rd positions respectively because there is no real threat entailed by their ability to establish CC when the likelihood of hope's effect being expressed is so low.

The placing of less greedy emotional characters in scenarios 2-5 (see tables 8.15, 8.16, 8.17 and 8.18) is identical to the placing of emotional characters in scenarios 2-5 when greed is not present within the population. The reasoning for this positioning pattern is exactly the same as outlined in my discussion of why the same positioning pattern occurs when no greed is present so I will not labour the point again.

Moderate Greed Since cooperation is initially ubiquitous in scenario 1, it would appear that the emotional characteristic that has the greatest premium placed upon it is tolerance when considering moderately greedy emotional character populations. If A2:G1 and A3:G1 are first compared using a standardised play history (as in table 8.19), it can be observed how A3:G1 is capable of delaying the establishment of DD that is not motivated by greed with its opponents until round 13 whereas DD that is not motivated by greed is established on round 9 for A2:G1 agents.

The higher placing of A3:G2 compared to A3:G1 in scenario 1 is due to pure chance. This was determined following an analysis of the play history of the agents that achieved the minimum scores responsible for the placings of A3:G1 and A3:G2. The focus of this chance is upon the *disposition* taken by each agent towards its opponents when hope is activated. The play histories in question are outlined in table 8.20 for reference and comparison. Plays that are made by agents under the influence of hope's effect are highlighted in bold in this table.

With respect to table 8.20, I draw attention to agent x in context of A3:G1 and A3:G2 since this is the agent which earns the score that is the subject of interest (1031 for

TABLE 8.19: How the increased tolerance of A3:G1 enables the maximisation of an individual's minimum score compared with the less tolerant A2:G1.

	A2:G1		**A3:G1**	
Round #	*Agent x*	*Agent y*	*Agent x*	*Agent y*
1	C	C	C	C
2	C	C	C	C
3	C	C	C	C
4	***D***	***D***	***D***	***D***
5	C	C	C	C
6	C	C	C	C
7	C	C	C	C
8	***D***	***D***	***D***	***D***
9	D	D	C	C
10	D	D	C	C
11	D	D	C	C
12	D	D	***D***	***D***
13	D	D	D	D
Individual Score	*25*	*25*	*31*	*31*

TABLE 8.20: How the probability of hope's effect being manifest in an agent after its activation can cause two similar emotional characters to achieve different minimum scores.

	A3:G1				**A3:G2**			
Round #	*Agent y*	*Agent x*	*Agent x*	*Agent z*	*Agent y*	*Agent x*	*Agent x*	*Agent z*
1	C	C	C	C	C	C	C	C
2	C	C	C	C	C	C	C	C
3	C	C	C	C	C	C	C	C
4	**D**	**C**	**D**	**D**	**D**	**D**	**D**	**D**
5	**D**	C	C	C	C	C	C	C
6	**D**	C	C	C	C	C	C	C
7	**D**	D	C	C	C	C	C	C
8	C	D	**D**	**D**	**D**	**D**	**D**	**D**
9	C	C	C	C	C	C	C	C
10	C	C	C	C	C	C	C	C
11	C	C	C	C	C	C	C	C
12	**C**	**C**	**C**	**C**	**D**	**C**	**C**	**D**
13	**D**	**C**	**C**	**D**	**D**	D	D	**D**
14	**D**	C	D	**D**	D	D	D	D
15	**D**	C	D	D	D	D	D	D
16	**D**	D	D	D	D	D	D	D
Individual Score	*53*	*28*	*35*	*40*	*38*	*33*	*33*	*38*

A3:G1, 1034 for emotional A3:G2). In round 4 all agents of both emotional characters have hope activated and are capable of periodically defecting in this round. Despite

this, whilst all A3:G2 agents defect against each other, agent x of A3:G1 cooperates and receives the sucker's pay-off in this round. Due to A3:G1's tolerant nature (a quality shared by A3:G2 agents), agent x must then incur 2 sucker's pay-offs before it may deactivate agent y's hope (greed) by defecting against it. Therefore, by the end of round 7 when A3:G1's agent x defects against its opponent y (deactivating y's hope) agent x has earned a total of 29 whereas agent x of A3:G2 has earned 38. This is due entirely to the A3:G2 agent x defecting against both its opponents in round 4 i.e. the difference in score emerges as a consequence of hope's probability of effect.

Still with reference to table 8.20, in round 12 it can be seen that all A3:G1 agents have hope activated but cooperate with their respective opponents. Due to this, all A3:G1 agents have hope activated again in round 13 yet, whilst agent x cooperates, its opponents defect. This causes the deactivation of x's hope (greed) and x reverts back to its base disposition of cooperation, receiving the sucker's pay-off 3 times from each opponent before finally establishing DD with both of its opponents, y and z. Likewise, by the end of round 13, both y and z A3:G2 agents have received their first 3 defections from agent x causing DD to be established between x and y, and x and z on round 14. Whilst DD is established with both opponents earlier for agent x of A3:G2, its re-establishment of CC plays, following the DDs triggered by hope's activation on rounds 4 and 8, maximises its score. By the conclusion of round 16, agent x of A3:G1 has earned 63 points whilst agent x of A3:G2 has earned 66 points.

The premium on tolerance for moderately greedy characters playing in the context of scenario 1 is also demonstrated by analysing the placings and play histories of the lowest scoring A1:G1, A2:G2 and A3:G3 agents. Table 8.21 makes the premium placed upon tolerance clearer:

TABLE 8.21: How different tolerance levels enable equally tolerant and responsive emotional characters to achieve different minimum system scores.

	A1:G1				**A2:G2**				**A3:G3**			
Round #	*Ag. y*	*Ag. x*	*Ag. x*	*Ag. z*	*Ag. y*	*Ag. x*	*Ag. x*	*Ag. z*	*Ag. y*	*Ag. x*	*Ag. x*	*Ag. z*
1	C	C	C	C	C	C	C	C	C	C	C	C
2	C	C	C	C	C	C	C	C	C	C	C	C
3	C	C	C	C	C	C	C	C	C	C	C	C
4	**D**	**C**	**C**	**D**	**D**	**C**	**C**	**D**	**D**	**C**	**C**	**D**
5	**D**	D	D	**D**	**D**	C	C	**D**	**D**	C	C	**D**
6	D	D	D	D	**D**	D	D	**D**	**D**	C	C	**D**
7	D	D	D	D	C	D	D	C	**D**	D	D	**D**
8	D	D	D	D	D	D	D	D	C	D	D	C
9	D	D	D	D	D	D	D	D	C	D	D	C
10	D	D	D	D	D	D	D	D	D	D	D	D
Individual Score	*20*	*15*	*15*	*20*	*23*	*18*	*18*	*23*	*26*	*21*	*21*	*26*

Table 8.21 shows the round number in which the lowest scoring agent of each emotional character, x, locks into DD. Note how this round number increases from round 6

for A1:G1 to round 8 for A2:G2, and to round 10 for A3:G3. This delayed establishment of DD occurs due to increased tolerance. If A1:G1 and A2:G2 are used as an example, it can be observed that, when hope is activated for all agents in round 4, agents y and z both defect whilst x cooperates. Agent x of A1:G1 then defects against y and z in round 5 causing their hope (greed) to be deactivated; this, coupled with the intolerance of y and z, causes DD to be established on round 6 between x and its opponents. Agent x of A2:G2 does not immediately punish defection with defection on round 5, instead it continues to cooperate until the end of round 6 when two defections have been received from its opponents causing a switch in disposition. Now, the tolerance of agents y and z becomes important because this further delays the establishment of DD until round 8 since agents y and z must also receive two defections from agent x before they punish defection with defection. This delay in DD being established maximises the individual score of agent x for A2:G2 when compared to agent x of A1:G1. Explaining the score of agent x of A3:G3 follows a similar pattern of reasoning so I will not elaborate further.

The placing of A2:G1 and A3:G3 is interesting with respect to scenario 1 since they have no similar emotional characteristics (A3:G3's tolerance and responsiveness ratio is equal whilst A2:G1's is not) yet they still manage to achieve the same minimum score (1022) in certain games. Therefore, the equal scoring observed can only be due to the probability of hope's effect being manifested in an agent after its activation. The play history detailed below in table 8.22 clarifies this point.

TABLE 8.22: How the probability of hope's effect being manifested in an agent after its activation can cause two different emotional characters to achieve the same minimum score.

	A2:G1				**A3:G3**			
Round #	*Agent y*	*Agent x*	*Agent x*	*Agent z*	*Agent y*	*Agent x*	*Agent x*	*Agent z*
1	C	C	C	C	C	C	C	C
2	C	C	C	C	C	C	C	C
3	C	C	C	C	C	C	C	C
4	***D***	***D***	***D***	***D***	***D***	***C***	***C***	***D***
5	C	C	C	C	***D***	C	C	***D***
6	C	C	C	C	***D***	C	C	***D***
7	C	C	C	C	***D***	D	D	***D***
8	***D***	***C***	***C***	***D***	C	D	D	C
9	***D***	D	D	***D***	C	D	D	C
10	D	D	D	D	D	D	D	D
Individual Score	*26*	*21*	*21*	*26*	*26*	*21*	*21*	*26*

The agent of interest in table 8.22 is agent x (for each emotional character type considered). With reference to all agents of all emotional character types, DD is locked into by round 10 yet the play histories of intermediate rounds are quite different. With reference to A2:G1, all agents have hope activated in round 4 due to ubiquitous initial cooperation yet, all agents defect, causing hope to be deactivated for all agents by the

conclusion of round 4. With regard to A3:G3 agents, in round 4 agent x cooperates with agents y and z following hope's activation however, y and z both defect against x causing the deactivation of x's hope (greed). Therefore, whereas agent x of A2:G1 enjoys 3 rounds of CC in rounds 5-7, agent x of A3:G3 suffers two sucker's pay-offs in rounds 5 and 6 followed by DD on round 7. Consequently, by the end of round 7, agent x of A2:G1 has earned 38 points whereas agent x of A3:G3 has earned 20 points. Despite this, agent x of A3:G3 increases its individual score by virtue of its unresponsiveness in rounds 8 and 9. Due to a combined effect of deactivating y and z's hope (greed) in round 7 and the activation of x's anger, x will continue to defect until 3 cooperations are received from an opponent. Since y and z's hope (greed) has been deactivated, their disposition reverts back to cooperation with agent x in rounds 8 and 9 as neither y or z has received three defections thus far from agent x. So, in rounds 8 and 9, agent x of A3:G3 earns 20 points by distributing the sucker's pay-off to both its opponents, increasing its individual score to 40 by round 9 and equalling the individual score obtained by agent x of A2:G1.

The value of tolerance is also demonstrated by the placement of A2:G2 and A2:G3 who are as tolerant as each other but place above the most intolerant emotional characters and below the most tolerant emotional characters. The placement of A1:G2, A1:G3 and A2:G3 in scenario 1 is to be expected for reasons already detailed (their propensity to establish DD after receiving 1 or 2 sucker's pay-offs). A2:G3 places slightly higher because it delays the establishment of DD plays as previously shown in table 8.6.

Finally, with respect to scenarios 2-5 (see tables 8.15, 8.16, 8.17 and 8.18) the same pattern of emotional character placings identified when greed is not/less likely to occur is observed. The reasoning behind this positioning pattern is exactly as described in the discussion of why this pattern emerges in non-greedy emotional character populations so will not be reiterated here.

High Greed The premium placed upon tolerance with regards to maximising the minimum score obtained is at its highest when the chance of defection occurring after greed has been triggered is highly likely. Therefore, with respect to scenario 1 (see table 8.14), emotional character placement and explanation thereof is identical to the placements/explanations outlined for when the level of greed in the population is moderate.

Again, the patterns of emotional character placing replicate those observed for non/-less/moderately greedy emotional character populations (see tables 8.15, 8.16, 8.17 and 8.18). The explanations behind these patterns is provided in my discussion of why non-greedy emotional characters place in this way.

8.5.1.3 Total System Score

Results Generally speaking, to ensure the maximisation of a system's total score, pairs of individuals within the population should establish CC as quickly as possible. CC between agent pairs will earn the system 6 points per round whereas CD/DC play earns the system 5 points per round and DD only earns the system 2 points per round.

Tables 8.23, 8.24, 8.25, 8.26 and 8.27 show the placing of emotional characters (organised by greed likelihood) with respect to scenarios 1-5. Again, as with table 8.7 in section 8.5.1.1, table 8.23 contains no column for emotional characters whose likelihood of greed is non-existent. This is because all individuals in these emotional character populations score 1014000 since all individuals cooperate with their opponents initially. This ultimately means that hope is never triggered, so periodic defection never has a chance to occur; in the context of this scenario and for emotional characters who will never act greedily, CC plays never break once established.

TABLE 8.23: Emotional character placing based upon maximum total system scores obtained in scenario 1 organised by the likelihood of hope's effect being manifest after activation.

Position	**H:1**	**H:2**	**H:3**
1	A1:G1	A1:G1	A3:G1
2	A3:G1	A2:G2	A3:G2
3	A2:G2	A3:G3	A3:G3
4	A2:G1	A3:G1	A2:G1
5	A3:G3	A2:G1	A2:G2
6	A3:G2	A3:G2	A1:G1
7	A2:G3	A2:G3	A2:G3
8	A1:G2	A1:G2	A1:G3
9	A1:G3	A1:G3	A1:G2

TABLE 8.24: Emotional character placing based upon maximum total system scores obtained in scenario 2 organised by the likelihood of hope's effect being manifest after activation.

Position	**H:0**	**H:1**	**H:2**	**H:3**
1	A3:G1	A1:G1	A1:G1	A3:G3
2	A3:G2	A3:G1	A2:G2	A2:G2
3	A2:G1	A2:G2	A3:G3	A1:G1
4	A1:G1	A3:G3	A3:G1	A3:G1
5	A3:G3	A2:G1	A2:G1	A3:G2
6	A2:G2	A3:G2	A3:G2	A2:G1
7	A2:G3	A2:G3	A2:G3	A2:G3
8	A1:G2	A1:G2	A1:G2	A1:G2
9	A1:G3	A1:G3	A1:G3	A1:G3

No Greed In context of scenarios 2-5 (see tables 8.24, 8.25, 8.26 and 8.27), A2:G1, A3:G1 and A3:G2 place highest, as expected. This occurs because these emotional characters establish CC from CD/DC plays in the fewest rounds (i.e. those that are more responsive than tolerant) enabling them to maximise total system scores most proficiently. As shown by table 8.4 in the introduction to section 8.5, A2:G1 and A3:G1 are the quickest to establish CC (one round after CD/DC has been played) and A3:G2 establishes CC after two rounds of CD/DC play.

TABLE 8.25: Emotional character placing based upon maximum total system scores obtained in scenario 3 organised by the likelihood of hope's effect being manifest after activation.

Position	**H:0**	**H:1**	**H:2**	**H:3**
1	A3:G2	A1:G1	A1:G1	A1:G1
2	A2:G1	A3:G1	A2:G2	A2:G2
3	A3:G1	A2:G2	A3:G3	A3:G3
4	A2:G2	A3:G3	A3:G1	A3:G1
5	A1:G1	A2:G1	A2:G1	A3:G2
6	A3:G3	A3:G2	A3:G2	A2:G1
7	A1:G2	A2:G3	A2:G3	A2:G3
8	A2:G3	A1:G3	A1:G3	A1:G3
9	A1:G3	A1:G2	A1:G2	A1:G2

TABLE 8.26: Emotional character placing based upon maximum total system scores obtained in scenario 4 organised by the likelihood of hope's effect being manifest after activation.

Position	**H:0**	**H:1**	**H:2**	**H:3**
1	A3:G2	A1:G1	A3:G3	A1:G1
2	A2:G1, A3:G1	A2:G2	A1:G1	A2:G2
3	A1:G1	A3:G3	A2:G2	A3:G3
4	A2:G2, A3:G3	A3:G1	A3:G1	A3:G1
5	A1:G2	A2:G1	A2:G1	A3:G2
6	A1:G3	A3:G2	A3:G2	A2:G1
7	A2:G3	A2:G3	A2:G3	A2:G3
8		A1:G2	A1:G3	A1:G3
9		A1:G3	A1:G2	A1:G2

TABLE 8.27: Emotional character placing based upon maximum total system scores obtained in scenario 5 organised by the likelihood of hope's effect being manifest after activation.

Position	**H:0**	**H:1**	**H:2**	**H:3**
1	A3:G1	A1:G1	A2:G2	A2:G2
2	A2:G1	A3:G3	A3:G3	A3:G3
3	A3:G2	A2:G2	A1:G1	A1:G1
4	A1:G1	A3:G1	A3:G1	A3:G1
5	A2:G2	A2:G1	A2:G1	A3:G2
6	A3:G3	A3:G2	A3:G2	A2:G1
7	A2:G3	A2:G3	A2:G3	A2:G3
8	A1:G3	A1:G3	A1:G2	A1:G2
9	A1:G2	A1:G2	A1:G3	A1:G3

The placing of A1:G1, A2:G2 and A3:G3 with respect to scenarios 2-5 is also expected since these emotional characters are incapable of converting CD/DC plays into CC plays; instead they maintain CD/DC plays (see table 8.5). This earns the system 5 points every round instead of 6 as with CC.

A1:G2, A1:G3 and A2:G3 place lowest because CD/DC plays are *always* converted into DD (see table 8.6). DD is then maintained for the remainder of the game, earning the total system a score of around 338000 (2 points per round for each pair of agent opponents).

Less Greed With respect to scenarios 1-3 (see tables 8.23, 8.24 and 8.25), A1:G1, A3:G1 and A2:G2 place first, second and third respectively. The success of A3:G1 is expected because if CD/DC plays should occur initially or after greed is triggered, an agent with this emotional character is capable of establishing CC after one round. The placing of A1:G1 and A2:G2 is interesting since one might expect A2:G1 and A3:G2 to place alongside A3:G1 in the top 3 positions since they are also capable of establishing CC after a CD/DC play has occurred (A3:G2 achieves this albeit one round later than A2:G1 and A3:G1).

The success of A1:G1 and A2:G2 in scenarios 1-3 provides evidence that when greed is introduced to the system, there is a premium placed upon emotional characters that *avoid* activating hope i.e. to avoid establishing CC, because hope makes it possible for DD to be established between agents[2]. Therefore, the placing of A1:G1 and A2:G2 is due to their CD/DC play maintenance. This also explains why A3:G3 places above A2:G1 and A3:G2 since its play style mimics the play style of A1:G1 and A2:G2. Whilst TFT style play does not earn an optimal amount for the system (5 points per individual, per round instead of 6), it earns more than the lowest amount of points available (2 points per individual, per round). Therefore, it can be asserted that: by increasing the level of greed present within the population, a premium is placed upon emotional characters that avoid the establishment of DD and CC when the total system score is considered.

Scenarios 4 and 5 (see tables 8.26 and 8.27) demonstrate the extent of the premium placed upon the avoidance of converting CD/DC plays into CC (which can, given greed, lead to DD plays being established) since A1:G1, A2:G2 and A3:G3 place in the top three positions and A2:G1, A3:G1 and A3:G2 place immediately below.

With regards to all scenarios, A1:G2, A1:G3 and A2:G3 (all of which are less tolerant than responsive) are again least successful since CD/DC plays are always converted into DD plays. Also, due to the added chance of DC/CD plays occurring initially as scenario number increases and the inclusion of CD/DC occurrence due to the manifestation of hope's effect in agents, such emotional characters perform poorly with respect to total system score.

Moderate Greed Moderately greedy emotional characters have a 0.5 chance of defecting after hope is activated, giving rise to the play scenarios outlined in table 8.3.

[2]This causes the total system score to suffer. Indeed, the total system score always suffers if one agent should ever destabilise CC but less so if CD/DC plays are established and maintained rather than DD

Therefore, because all four play scenarios are equally likely to occur, there is an increased premium on emotional characters that can avoid establishing CC (which activates hope) and DD (which minimises the total system score) i.e. those whose tolerance and responsiveness ratios are equal. Due to this, A1:G1, A2:G2 and A3:G3 consistently place in the top 3 positions since they maintain CD/DC plays after establishment whilst A2:G1, A3:G1 and A3:G2 place in positions 4-6 due to their ability to establish CC from CD/DC plays. A1:G2, A1:G3 and A2:G3 are the least successful at maximising total system score in all scenarios due to their characteristic DD establishment following CD/DC plays.

High Greed Results from scenario 1 (see table 8.23) see A3:G1, A3:G2 and A3:G3 placing highest whilst A1:G1, A2:G1 and A2:G2 place below. The placing of these emotional characters is interesting since it would appear that the premium on increased tolerance and responsiveness is high. This is reasonable since all agents in scenario 1 have hope activated by the conclusion of round 3 due to ubiquitous initial cooperation. When this situation is combined with the fact that the likelihood of establishing DD is so high on round 4 (due to the high likelihood of agents acting greedily) then the reasoning behind why maximally tolerant agents are selected begins to emerge.

To maximise total system scores, agents must either re-establish CC (earning the system 6 points per round) or maintain CD/DC plays (earning the system 5 points per round) that may be established in round 4. If a situation is considered where an agent x and its opponent y have hope activated in round 4 and x cooperates whilst y defects, then it can be seen how tolerance is able to maximise total system score. In this situation, if x is highly tolerant, it will continue to cooperate in the face of defection from y for up to 3 rounds. This cooperation from x causes y's hope to be continually activated in these 3 rounds, earning the system a total of 15 points. If x is quicker to meet defection with defection due to moderate tolerance then the system will only earn a total of 10 points. Increased tolerance therefore enables the greater sacrifice of individual score for the benefit of the system, a feature of tolerance first described in chapter 6. In other words: it is better from the standpoint of the total system for an individual to receive the sucker's pay-off than it is to establish DD. From a socialist perspective, this observation has parallels with taxation, where people incur an individual loss for the greater good of the system.

The placing of A2:G1, A3:G1 and A3:G2 in the context of all scenarios serves to demonstrate how responsiveness also affects maximisation of total system scores. From the standpoint of the *total system*, establishment of CC is still preferred even under such conditions where agents are very likely to defect periodically following hope's activation. Maintenance of CD/DC is the most preferred outcome in reality but the system always wants the total score to be maximised as much as possible. Under highly greedy conditions a combination of CC and maintaining CD/DC plays is much preferred indeed, any play other than DD is preferred. To clarify, consider two agents, x and y, that are

acting under the influence of hope (greed). Under this influence, both x and y defect, deactivating hope (greed) in both agents. From here, x and y may either re-establish CC (since x and y *must* have been cooperating or hope would not have been activated), a CD/DC play may occur or DD may be established (depending upon what has occurred before the DD motivated by hope/greed occurs).

If a CD/DC play is established between x and y then increased responsiveness increases the likelihood that CC will be re-established rather than DD. Increased responsiveness is welcomed from the standpoint of the system because two rounds of CD/DC followed by three rounds of CC earns the system 28 rather than 23 (obtained by one round of CD/DC followed by three rounds of CC), but responsiveness can only be increased to a point. As argued for in chapter 7, the moderate responsiveness of A3:G2 has a "goldilocks" level of responsiveness (it is neither too responsive nor too unresponsive, its responsiveness is just right) enabling two rounds of CD/DC play to occur before the re-establishment of CC. Conversely its reduced responsiveness (compared to A3:G1) means that it runs a greater risk of establishing DD. Hence A3:G1 always places above A3:G2 due to its "safe" exploitation.

As the likelihood of establishing DD increases, maintenance of CD/DC plays becomes more important. If scenarios 2-5 are considered (see tables 8.24, 8.25, 8.26 and 8.27) A1:G1, A2:G2 and A3:G3 place in the top three positions. The ability of these emotional characters to maintain CD/DC plays after their establishment (see table 8.5) is extremely beneficial since this avoids the potential for DD to be established from an establishment of CC.

A2:G1, A3:G1 and A3:G2 place below A1:G1, A2:G2 and A3:G3 in the context of scenarios 2-5 due to their ability to establish CC from DC/CD plays (see table 8.4) which increases the likelihood of DD being established through hope (greed).

Finally, A1:G2, A1:G3 and A2:G3 place below A2:G1, A3:G1 and A3:G2 in scenarios 2-5 because they always convert CD/DC plays into DD which earns the total system the least amount of points possible (see table 8.6).

8.5.2 Behavioural Features of Emotional Characters with Differing Tolerance and Responsiveness Ratios

In section 8.5.1, emotional characters were grouped and evaluated according to their tolerance and responsiveness ratios. This gives rise to an interesting topic of research concerning the behavioural features of emotional characters with different tolerance and responsiveness ratios (see research question 6 in section 8.1). The intention of this section is to therefore provide an answer to this research question by first identifying a subset of emotional characters to use and then distinguishing some variables of interest to analyse (in much the same way as I did in section 8.5.1 with respect to maximum/minimum individual scores and total system score).

The tolerance and responsiveness ratio groupings mentioned above were described in the introduction to section 8.5. From these three groupings, three individual emotional

characters were selected for investigation, namely A2:G1, A2:G2 and A2:G3 along with their greedy counterparts. These emotional characters were chosen since their tolerance is equal whereas their responsiveness differs and it is the relation between tolerance and responsiveness that matters (see B). This ensures the fairest and easiest comparison possible of the three groupings outlined above. Furthermore, these emotional characters span the three groupings mentioned (more responsive than tolerant; equally tolerant and responsive; and less tolerant than responsive). A number of different criteria were identified so that any differences between behavioural features for the emotional characters considered can be differentiated and analysed, these are listed below.

- Average number of distinct scores for non-greedy and greedy variants of A2:G1, A2:G2 and A2:G3 populations.
- Average percentage of each non-greedy and greedy variant of A2:G1, A2:G2 and A2:G3 populations scoring:
  - > 3000
  - 3000
  - Between and including 2999 - 1001
  - 1000
  - < 1000
- The equality of each non-greedy and greedy variant of A2:G1, A2:G2 and A2:G3 populations.

As such, this section is divided into three subsections each one dealing with one of the variables of interest presented above with respect to populations of A2:G1, A2:G2 and A2:G3. Section 8.5.2.1 is concerned with an analysis of the average number of distinct scores; section 8.5.2.2 provides an analysis of the average percentage of each emotional character population considered scoring within each of the five brackets identified above; section 8.5.2.3 deals with the equality of each of the emotional character populations considered and their greedy variants. Each section is introduced with an explanation of *why* the measure discussed was selected and some further clarification regarding the measure in question.

8.5.2.1 Distinct Score Analysis

Introduction The average number of distinct scores present within each emotional character population facilitates commentary upon what degree of variation in behaviour exists within the emotional character population considered; the lower the number of distinct scores the lower the score variation within that population. Less variation indicates that the amount of unprovoked defection occurring due to the temptation to be greedy is stifled whilst high variation indicates the opposite. From this I will then

be able to judge the extent to which each emotional character population is capable of establishing societal norms.

The decision to calculate average numbers of distinct scores for each character population was taken since, due to the 5 repeats performed for each emotional character population, some variation in this criterion exists. Rather than selecting the maximum or minimum number of distinct scores, which may not be a completely true measure (the emotional character population may be capable of generating more or less distinct scores), I take an average measurement so that a general explanation for the emotional character population can be provided.

Note that the results discussed in this section do not mention the average number of distinct scores achieved in context of scenario 6. This is because DD is ubiquitous initially in this scenario thus, populations only ever contain one distinct score.

No Greed In analysing the average number of distinct scores for each non-greedy population of A2:G1, A2:G2 and A2:G3, I have given no attention to scenario 1 (as well as scenario 6 as previously explained). Since cooperation is initially ubiquitous in the context of this scenario and since non-greedy agents never periodically defect, the average number of distinct scores achieved by non-greedy A2:G1, A2:G2 and A2:G3 populations are always equal (1). Consequently, tables containing average numbers of distinct scores for the non-greedy A2:G1, A2:G2 and A2:G3 populations do not contain a row for scenario 1.

If the non-greedy A2:G1 population is first considered there are always six distinct scores achieved irrespective of scenario. These distinct scores are listed below together with an explanation of the behaviour that achieves the score in question:

- 1000 - An agent establishes DD with its opponents on the initial round and maintains this for the remainder of the game.
- 2002 - An agent, x, that initially defects is met by one initially cooperative opponent, y, and one opponent that initially defects, z. By the conclusion of round 1, x's gratitude potential towards y has reached the emotion's activation threshold and subsequently, CC between x and y is established on round 2 and maintained for the remainder of the game. With respect to agents x and z, DD is established on the initial round and maintained for the remainder of the game.
- 2994 - An agent, x, receives the sucker's pay-off from both of its opponents on the initial round followed by the establishment of CC with both opponents in round 2 since x's cooperation activates gratitude in the opponents. These CCs are then maintained for the remainder of the game.
- 2997 - An agent, x, receives the sucker's pay-off from one of its opponents, y, and cooperation from its other opponent, z, on the initial round. Since x cooperates with y gratitude is activated in y and CC is established between x and y on round

2. The CC initially established between x and z is then maintained along with x and y new-found CC for the remainder of the game.

- 3000 - An agent establishes CC with both of its opponents on the initial round and maintains it for the remainder of the game.

- 3004 - An agent, x, initially defects against both its opponents, y and z, who cooperate with x in the initial round. In round 2, x cooperates with both y and z since its gratitude has been activated. This establishes CC between agent x and agents y and z on round 2 and these CCs are maintained for the remainder of the game.

Like the non-greedy A2:G1 population, the non-greedy emotional A2:G3 population always produces six individual scores for scenarios 2-5. Detailed below are the individual scores and the behaviours which produce them:

- 996 - An initially cooperative agent, x, is met with defection from both opponents y and z in round 1. Agent x is then exploited for two rounds until its anger is activated whereupon it establishes DD with y and z in round 3 and maintains these plays for the remainder of the game.

- 1000 - An agent establishes DD with both of its opponents on the initially and maintains DD for the remainder of the game.

- 1008 - Occurs when an agent, x, initially defects and is met with initial cooperation from an opponent, y, and initial defection its other opponent, z. DD is then established by x and z in round 1 and maintained for the remainder of the game. However, x is able to exploit y for two rounds before y's anger is activated and DD is established between them. This DD is then maintained by x and y for the remainder of the game along with x and z's DD causing the score observed to be achieved.

- 1016 - This score is achieved in a similar way to the way a score of 1008 is achieved however, both y and z cooperate initially allowing x to distribute the sucker's pay-off for two rounds before establishing DD with both y and z and maintaining these plays for the remainder of the game.

- 1998 - Achieved when an initially cooperative agent, x, is met with defection from one opponent, y and cooperation from the other, z. CC is then established in round 1 between x and z and maintained for the remainder of the game. However, y is able to exploit x for two rounds before DD is established between these agents on round 3 and maintained for the remainder of the game.

- 3000 - An agent establishes CC with its opponents initially round and maintains these plays for the remainder of the game.

The scores and behaviours observed in the non-greedy A2:G3 population mirror the scores and behaviours observed in the non-greedy emotional A2:G1 population i.e. initial defectors who exploit their opponents lock into DD plays rather than CC plays due to their reduced responsiveness. As explained in the introduction to section 8.5, emotional characters whose tolerance and responsiveness ratios are weighted in favour of unresponsiveness (such as A2:G3) always establish DD following CD/DC plays (see table 8.6). Conversely, emotional characters whose tolerance and responsiveness ratios are weighted in favour of tolerance (such as A2:G1) always establish CC following CD/DC plays (see table 8.4). The behaviour of the non-greedy variant of A2:G1 results in higher distinct scores scores being achieved when compared against the non-greedy variant of A2:G3.

Compared with the non-greedy populations of emotional characters A2:G1 and A2:G3, there are fewer distinct scores achieved by the non-greedy A2:G2 population. Instead of 6 distinct scores, only 5 are achieved due to A2:G2's TFT (specifically "two-tit-for-two-tat") behaviour. Such behaviour ensures that agents are never able to exploit an opponent before locking into CC or DD. The scores achieved by the A2:G2 population are listed below along with an explanation of how such a score is achieved:

- 1000 - An agent establishes DD with both its opponents on the initial round and maintains this for the remainder of the game.
- 1750 - Occurs when an agent, x, initially defects against an initially cooperative opponent, y, and another initial defector, z. In round 2, z defects again, activating x's anger and resulting in DD being completely established between x and z on round 3 and maintained for the remainder of the game. With respect to agent y however, on round 3, x and y switch disposition i.e. x cooperates with y and y defects against x. This occurs because their anger and gratitude emotions have been activated and this two-tit-for-two-tat play is maintained for the remainder of the game.
- 2500 - Obtained when an agent establishes a two-tit-for-two-tat play as described above with *both* of its opponents in the initial round. This play is then maintained for the remainder of the game.
- 2750 - Occurs when an initially cooperative agent, x, is pitted against an initial defector, y, and an initial cooperator, z. In round 3, x and z reciprocate each others involuntary cooperation by virtue of gratitude and CC is established and maintained between x and z for the remainder of the game. However, on round 3, x and y switch tactics and this two-tit-for-two-tat play is maintained for the remainder of the game.
- 3000 - An agent establishes CC with its opponents initially and maintains these plays for the remainder of the game.

Greed The introduction of greed into A2:G1, A2:G2 and A2:G3 populations causes the number of distinct scores present within each population to increase dramatically with the exception of A2:G2 population variants whose average number of distinct scores are still restrained. Therefore, specific discussion of exact distinct scores and how they are achieved for greedy populations of A2:G1 and A2:G3 will no longer be offered. Instead, my discussion of these emotional character populations will concentrate upon general trends observed since this is both significantly more fruitful and less time consuming.

Unlike the discussion of non-greedy variants of A2:G1, A2:G2 and A2:G3 above, the average number of distinct scores for scenario 1 are now considered since periodic defection is now possible. This results in more than one distinct score being achieved for these emotional character populations and therefore some interesting comments may be passed. Bear in mind that there are 338 agents in total in each emotional character population and average numbers of distinct scores should be referenced against this number.

TABLE 8.28: Average number of distinct scores for greedy A2:G1, A2:G2 and A2:G3 populations in context of scenarios 1-5.

		Avg. # Distinct Scores Per Pop.		
Hope Character	**Scenario #**	*A2:G1*	*A2:G2*	*A2:G3*
H:1	1	300	133	90
	2	294	113	91
	3	286	80	66
	4	238	52	45
	5	156	19	20
H:2	1	180	57	24
	2	184	55	35
	3	177	44	32
	4	151	31	26
	5	107	15	14
H:3	1	88	19	14
	2	154	24	23
	3	175	22	21
	4	157	19	17
	5	115	9	11

If the average number of distinct scores obtained is taken into consideration for all greedy variants of each emotional character population considered (A2:G1, A2:G2 and A2:G3) then it can be asserted that A2:G1 enables a greater variation of distinct scores. However, as it becomes harder to elicit gratitude from other agents, there is an observed reduction in the number of distinct scores obtained. As can be seen in table 8.28, A2:G3 populations contain a much reduced number of distinct scores for each scenario when compared against the average number of distinct scores obtained in each scenario by A2:G1 and A2:G2.

The results from table 8.28 illustrates how the numbers of average distinct scores differ between greedy variants of A2:G1, A2:G2 and A2:G3 populations. At first sight, it can be reasonably inferred that A2:G1 permits the greatest number of distinct scores to be achieved compared to A2:G2 and A2:G3 and A2:G2 permits a greater number of distinct scores to be achieved than A2:G3. However, in order to confirm this I have also calculated the average number of average distinct scores obtained for all greedy variants of A2:G1, A2:G2 and A2:G3 populations. These values are also presented in table 8.29 and clearly show that A2:G1 enables the greater number of distinct scores to be obtained in context of all greed likelihoods followed by A2:G2. As expected, A2:G3 achieves the lowest number of distinct scores.

If the average distinct scores for each greedy emotional character variant over the five scenarios are used, then the standard deviations of their average distinct scores can be determined. This in turn allows me to comment upon how much each emotional character restricts variations in plays for agents. The analysis performed yields table 8.29 which clearly illustrates the extent that A2:G3 restricts the number of average individual scores obtained. Therefore, it can be posited that emotional characters who are quicker to show anger than to show gratitude obtain *less* varied distinct scores in their populations. Conversely, emotional characters who are quicker to show gratitude than to show anger obtain *more* varied distinct scores in their population. Ergo, in this context, punishment encourages conformity; tolerance fosters diversity.

TABLE 8.29: Averages and standard deviations for number of distinct scores obtained over scenarios 1-5 for greedy variants of A2:G1, A2:G2 and A2:G3 populations.

	Avg. # (Std. Dev.) of Avg. Distinct Scores		
Hope Character	*A2:G1*	*A2:G2*	*A2:G3*
H:1	255 (60)	79 (46)	62 (31)
H:2	160 (32)	40 (17)	26 (8)
H:3	138 (35)	18 (6)	17 (5)

General Discussion The standard deviations and the average numbers of average distinct scores obtained for greedy variants of A2:G1, A2:G2 and A2:G3 (found in table 8.29) are obtained due to the difference in responsiveness and tolerance ratios.

As explained in the introduction to section 8.5 and table 8.4, A2:G1's activation threshold for gratitude being set lower than its activation threshold for anger enables this emotional character to convert CD/DC plays into CC. Establishing CC ensures that hope will be activated in the future giving an agent some chance to randomly defect or cooperate. The uncertainty of cooperation or defection that follows hope's activation enables various different plays to occur at various different times. If hope is activated more frequently this results in the number of distinct individual scores and their standard deviations being greater on average. Hence the observed increase in the values obtained

for these measures for A2:G1 when compared to the values obtained for these measures by A2:G2 and A2:G3.

The activation thresholds for gratitude and responsiveness for A2:G2 on the other hand are equal. This causes the emotional character to maintain CD/DC plays when they are established reducing the chance of hope being reactivated and reducing the chance of other play styles occurring (see table 8.5).

The tendency of A2:A3 to convert CD/DC plays into DD (see table 8.6) has been well documented throughout this chapter and its influence is exerted in this context also. The only way A2:G3 can establish CC is through pure chance, either with respect to initial disposition configurations or hope's activation and effect. For this emotional character, CC will never be achieved from CD/DC plays.

With respect to all emotional characters, table 8.29 illustrates the effect of greed likelihood upon the number of distinct scores and their concentration. As the likelihood of greed is increased, populations see a decrease in both the number of distinct scores and their spread. The explanation for this is relatively simple: as greed becomes more likely, there is more of a chance that DD will be established when hope is activated in agents. DD cannot be broken unlike CC in these simulations and as the likelihood of their occurrence is increased the number of distinct scores and their spread is reduced.

The implications of score variation are discussed further in sections 8.5.2.2 and 8.5.2.3 respectively but the foundations for explaining these implications have been laid in this section. From here a discussion of population percentages achieving distinct scores and fairness ratings can be calculated.

8.5.2.2 Percentage of Population Achieving Distinct Scores Analysis

Introduction In discussing distinct scores and their numbers within each emotional character population, commenting upon what percentage of the population is achieving certain scores is unavoidable. Not performing this task could result in misleading conclusions about an emotional character being drawn. For example: it could be said that the non-greedy variant of A2:G1 is best at maximising individual scores since it contains members of the population that score 3000 and above. However, if only a small percentage of the population achieves these scores whilst a higher percentage score only 1000 then claiming that this emotional character is adept at maximising individual scores would be incorrect.

In this section my intention is to pass comment upon percentages of A2:G1, A2:G2 and A2:G3 populations that fall into the 5 score brackets outlined in the introduction to section 8.5.2. These score brackets enable me to provide a number of interesting remarks regarding the prevalence of certain behaviours in each emotional character population considered. The score brackets that are considered and what information they provide about behaviours in the population are listed below.

- Scores >3000 indicate CC between agents for most of the game although some agents in the population have exploited their opponents for a number of rounds before CC was established and maintained.
- Scores equal to 3000 indicate CC between agents and their opponents for the whole game.
- Scores ≤2999 and ≥1001 indicate that agents have neither mutually cooperated or defected with or against their opponents for an entire game. Rather, some variation in behaviour exists.
- Scores equal to 1000 indicate that agents have mutually defected against both opponents for the entire game.
- Scores <1000 indicate that some agents have mutually defected with their opponents for most of a game however, some agents have been exploited by their opponents before DD was established and locked into.

Again, because there are five repeats run for each scenario, I have calculated the average number of agents that score within each of the brackets above to help scientifically sound and rigorous comparisons to be made. To perform this calculation the following method was implemented (listed for clarification). The values obtained for each emotional character population and its greedy variants are listed in tables 8.30, 8.31 and 8.32.

1. For each repeat of each scenario for each emotional character population, extract the final scores for each individual agent and calculate what percentage of these scores fall into each of the score brackets listed above.
2. Using these values for each of the five repeats of each scenario, calculate the average percentage achieving each score threshold for each scenario (except scenario 6 since this is the control scenario and all agents in all populations score 1000 due to ubiquitous initial defection).
3. Step 2 results in an average percentage for each of the five score brackets in each scenario for each emotional character population.

General Discussion I begin this discussion with a consideration of the non-greedy A2:G1 population. With reference to table 8.30, the average percentage of the population scoring within each bracket alters in an expected way:

- >3000 - A greater percentage of the population scores in this bracket, on average, as scenario number is increased towards scenario 3. It is then the case that there is a general trend for a reduced percentage of the population to score in this bracket as scenario number is increased from 3-5.

TABLE 8.30: Average percentage of all A2:G1 greedy variant populations that achieved scores within determined score brackets in context of scenarios 1-5.

		Avg. % Pop. Scoring *n*				
Hope Character	**Scenario #**	*>3000*	*3000*	*≤2999 and ≥1001*	*1000*	*<1000*
H:0	1	0	100	0	0	0
	2	12	51	36	1	0
	3	13	22	57	8	0
	4	11	6	61	22	0
	5	3	1	44	52	0
H:1	1	0	0	100	0	0
	2	0	0	99	1	0
	3	0	0	94	6	0
	4	0	0	78	22	0
	5	0	0	50	50	0
H:2	1	0	0	100	0	0
	2	0	0	99	1	0
	3	0	0	93	7	0
	4	0	0	78	22	0
	5	0	0	49	51	0
H:3	1	0	0	100	0	0
	2	0	0	98	2	0
	3	0	0	93	7	0
	4	0	0	78	22	0
	5	0	0	50	50	0

- 3000 - The percentage of the population, on average, obtaining this score decreases as more defectors are initially present within the population.
- ≤2999 and ≥1001 - The percentage of the population scoring within this bracket, on average, increases as scenario number is increased but then decreases from scenario 4 onwards.
- 1000 - The percentage of the population, on average, obtaining this score increases as more defectors are initially present within the population.
- <1000 - No agent in the population scores less than 1000.

These trends are expected due to the play probabilities incurred by the scenario played (see table 8.2) and the behavioural characteristics of A2:G1. Explanations for why these trends emerge are given below in context of the non-greedy A2:G1 population:

- >3000 - In scenario 1 all agents mutually cooperate resulting in CD/DC plays never being established. Due to this, no agent is capable of exploiting an opponent before establishing CC. However, as scenario number increases, the likelihood of CD/DC plays being established initially increases with initial CD/DC plays becoming most

TABLE 8.31: Average percentage of all A2:G2 greedy variant populations that achieved scores within determined score brackets in context of scenarios 1-5.

		Avg. % Pop. Scoring *n*				
Hope Character	**Scenario #**	*>3000*	*3000*	*≤2999 and ≥1001*	*1000*	*<1000*
H:0	1	0	100	0	0	0
	2	0	50	49	1	0
	3	0	21	72	6	0
	4	0	6	73	22	0
	5	0	1	49	50	0
H:1	1	0	0	100	0	0
	2	0	0	99	1	0
	3	0	0	94	6	0
	4	0	0	79	21	0
	5	0	0	49	51	0
H:2	1	0	0	100	0	0
	2	0	0	99	1	0
	3	0	0	94	6	0
	4	0	0	77	23	0
	5	0	0	50	50	0
H:3	1	0	0	100	0	0
	2	0	0	99	1	0
	3	0	0	94	6	0
	4	0	0	78	22	0
	5	0	0	49	51	0

likely in scenarios 3 and 4. Hence there is an increase in the percentage of the population achieving scores within this bracket since a greater number of initial defectors have a greater opportunity of suckering initial cooperators. The difference in percentage observed between scenarios 3 and 4 is most probably due to the fact that there are more initial defectors present in the population in the context of scenario 3 so exploitation of initial cooperators occurs more frequently. This situation reverses in scenario 4 however so more agents initially cooperate than defect so less agents score >3000. The drop in population percentage scoring within this bracket from scenario 4 to 5 is accounted for by the much reduced likelihood of CD/DC plays occurring initially and the increased prevalence of defectors initially meaning that there are less cooperators to exploit.

- 3000 - The general trend observed is simply accounted for by the fact that, as scenario number increases, the probability of meeting upon CC with opponents is decreased. Hence there is an inverse relationship between the percentage of the population scoring in this bracket and scenario number.

- ≤2999 and ≥1001 - As was explained with the >3000 score bracket, the likelihood of CD/DC plays occurring initially increases with scenario number and reaches its

TABLE 8.32: Average percentage of all A2:G3 greedy variant populations that achieved scores within determined score brackets in context of scenarios 1-5.

		Avg. % Pop. Scoring *n*				
Hope Character	**Scenario #**	*>3000*	*3000*	*≤2999 and ≥1001*	*1000*	*<1000*
H:0	1	0	100	0	0	0
	2	0	51	46	1	3
	3	0	22	61	7	10
	4	0	6	59	22	14
	5	0	1	35	52	12
H:1	1	0	0	100	0	0
	2	0	0	96	1	3
	3	0	0	84	6	10
	4	0	0	65	20	15
	5	0	0	36	52	12
H:2	1	0	0	100	0	0
	2	0	0	95	1	3
	3	0	0	84	6	10
	4	0	0	64	22	14
	5	0	0	36	51	12
H:3	1	0	0	100	0	0
	2	0	0	96	1	3
	3	0	0	84	6	10
	4	0	0	64	22	14
	5	0	0	36	51	12

peak in scenarios 3 and 4 before decreasing in scenario 5. This results in a greater concentration of individual scores within this bracket for these scenarios.

- 1000 - As with the explanation for the 3000 score bracket, the trend observed for this score bracket is accounted for by the increase of initial defectors as scenario number increases. This makes DD more likely to be established initially. Therefore, as scenario number increases, the percentage of the population scoring 1000 increases.

- <1000 - Since A2:G1 agents always convert initial CD/DC plays into CC, there is no chance of an agent being exploited before establishing DD. No agent therefore achieves scores within this bracket.

Most of these trends hold for non-greedy variants of A2:G2 and A2:G3 also, but with two notable exceptions. Firstly, A2:G2 and A2:G3 populations do not contain agents that score >3000 points because initial CD/DC plays are either maintained by A2:G2 agents or converted into DD by A2:G3 agents. Secondly, the only population that scores <1000 is A2:G3 since it is the only emotional character of those considered in this section capable of establishing DD following initial CD/DC plays. What is also interesting to note is the average percentages of the A2:G2 population scoring ≤2999 and

≥1001. When compared with the populations of A2:G1 and A2:G3, there is a greater percentage of the population (at least 5% more) achieving these scores in context of scenarios 2-5. This is due to the two-tits-for-two-tats behaviour possessed and expressed by A2:G2 which maintains CD/DC plays when they occur.

Average population percentages scoring within the brackets defined for greedy emotional character variants are markedly different when compared to their non-greedy counterparts. Populations of A2:G1 agents can no longer attain scores >3000 since CC causes a potential manifestation of hope's effect meaning that exploitation in initial rounds can never be followed with maintained CC. Furthermore, no emotional character population contains agents that score equal to 3000 for the same reason as just mentioned. Therefore, all greedy variants of the A2:G1 and A2:G3 populations see an increased percentage of their populations scoring within the ≤2999/ ≥1001 and 1000 brackets.

An inverse relationship exists with respect to average percentages of all greed variants of A2:G1 and A2:G2 populations scoring in the ≤2999/ ≥1001 and 1000 brackets. As scenario number is increased from 1 to 6, the percentage of these populations scoring ≤2999 and ≥1001 decreases and the percentage of agents scoring in the 1000 bracket increases. This is to be expected since the chance of two agents meeting upon DD initially increases as the initial percentage of defectors increases (see table 8.2). What is interesting however, is the fact that the two percentages equal out in scenario 5 for both emotional character populations. In addition, the standard deviation of the values that yield these averages always equals 1.17 for the A2:G1 population and 0.89 for the A2:G2 population. The smaller standard deviation for A2:G2 demonstrates how adept characters that employ TFT behaviour are at regulating and ensuring conformance to certain play styles within the population.

Greedy populations of A2:G3 still retain a percentage of the population that scores <1000. This is because there is no way of destabilising DD in the same way as CC: ergo, exploiting individuals before establishing DD for the rest of a game can still occur.

It is interesting that making greed more likely to occur (specifically from a 0.1 to 0.5 probability) has no noticeable effect upon population percentages scoring within the determined boundaries. Most notably with respect to this, one would have expected to see a general trend for all emotional characters where, as the temptation to be greedy increases, a higher percentage of the population scores in the 1000 bracket. The only reason why this should occur is due to the randomness of hope's effect which ensures that some agents score more than 1000 despite the likelihood of greed increasing.

8.5.2.3 Equality Analysis

Introduction Finally in this section, I turn to an analysis of equality for A2:G1, A2:G2 and A2:G3. To measure equality I calculate the *Gini coefficient*[3] for each of the

[3]Developed by Corrado Gini, the "Gini coefficient" is a commonly used measure of inequality amongst some frequency distribution, such as income. Details of how this value is calculated can be found in [72].

five repeats of every scenario for each relevant emotional character and then take an average of these values to produce an average Gini coefficient for each emotional character considered in every scenario. The results of these calculations for each emotional character population are listed in tables 8.33, 8.34 and 8.35. When a Gini coefficient is discussed in this section it is this table that should be used as a reference. A Gini coefficient of 0 indicates perfect equality in a population i.e. all agents have achieved the same individual score. The greater the Gini coefficient value, the more unequal the individual scores of agents within that population.

Since there are certain drawbacks to the relative nature of Gini coefficients e.g. two populations may have equal Gini coefficients but may differ greatly in wealth, I also consider the average median scores and average total system scores for each emotional character population in each scenario. These values are calculated in much the same way as the average Gini coefficient except that an average of the median and total score values for each of the five repeats of every scenario are taken. The values for each emotional character population considered are also listed in tables 8.33, 8.34 and 8.35 and these tables should be consulted when discussions of these values occur.

By using these three measures in conjunction I intend to provide a more rigorous discussion of equality since the total wealth of a character population is considered along with how equally that wealth is distributed amongst the population's agents. The main aim of this discourse is to identify trends within emotional character populations regarding wealth and its distribution and to classify each emotional character population in terms of some social order e.g. communism, capitalism etc. Note that scenario 6 is not considered in this discussion since average Gini coefficients, average median scores and average total scores are consistent for every character in the context of this scenario (0, 1000 and 338000 respectively). Consequently, Gini coefficients for scenario 6 do not appear in tables 8.33, 8.34 or 8.35.

No Greed If non-greedy emotional character variants are considered first it is observed that, depending upon scenario number, some emotional characters have greater degrees of equality than others. In scenarios 1 and 6, the equality of all non-greedy character populations are "perfect" (all individuals earn an equal amount) because CC/DD is never destabilised. With respect to all other scenarios, it is how CD/DC plays are dealt with that causes the scores detailed to be observed.

With regards to A2:G1, its non-greedy character population experiences a general increase in inequality as more defectors are introduced into the initial population. This increase is caused by the A2:G1's ability to convert CD/DC plays into CC: as the chance of DD occurring initially increases (as scenario number is increased, see table 8.2), more agents will lock into DD initially while a smaller percentage will initially encounter CD/DC plays. As shown in table 8.4, these plays will be converted into CC that will be maintained for the remainder of the game (as greed is not present so destabilisation cannot occur). Such plays earn the agents involved in the initial

Table 8.33: Average Gini coefficients, average median individual scores and average total system scores for all variants of A2:G1.

Hope Character	**Scenario**	**Avg. Gini Coefficient**	**Avg. Median Score**	**Avg. Total Score**
H:0	1	0.00	3000	1014000
	2	2.85	3000	983894
	3	10.15	2997	895846
	4	18.33	2795.6	774235.6
	5	24.95	1000	581092.8
H:1	1	12.84	1667.4	583881.4
	2	13.32	1660.9	578451.2
	3	14.89	1528.1	539934.8
	4	16.13	1339.8	487147.2
	5	14.41	1009.6	425836.2
H:2	1	3.76	1103.3	379630.6
	2	3.87	1106.3	380663.4
	3	4.02	1087.3	373892.2
	4	3.95	1062.6	365453.6
	5	3.07	1003.4	353101
H:3	1	2.98	1025.6	359462.2
	2	5.30	1064.1	375647.2
	3	5.87	1086.6	381909.6
	4	6.16	1067.2	377147.4
	5	4.80	1004	362158.8

CD/DC plays significantly more than those who establish and maintain DD from the initial round. Due to a combination of the following factors, the inequality of the A2:G1 population increases as more agents score 1000 and fewer agents score highly:

- A2:G1 is able to establish CC from CD/DC plays.
- More of the population establishes DD initially that is maintained for the remainder of the game.
- Less of the population meets upon CC and CD/DC initially.

The trend in equality for the non-greedy emotional A2:G2 population is similar to that of the non-greedy emotional A2:G1 population. Since agents of A2:G2 maintain CD/DC plays once established, as more initial defectors are introduced, a greater percentage of the population scores less whilst a smaller percent of the population scores more. The observed increase in inequality may then be accounted for.

The situation is not quite so simple for the A2:G3 population whose inequality rating increases up until scenario 3 where it then decreases as scenario number is increased. The inequality peak observed in scenario 3 is due to two factors: firstly, the probability of CD/DC plays occurring initially is at its peak in scenarios 3 and 4. Secondly, and perhaps more importantly, there is a higher probability of CC being met upon initially

TABLE 8.34: Average Gini coefficients, average median individual scores and average total system scores for all variants of A2:G2.

Hope Character	Scenario	Avg. Gini Coefficient	Avg. Median Score	Avg. Total Score
H:0	1	0.00	3000	1014000
	2	5.83	2900	933000
	3	11.90	2600	824600
	4	18.32	2200	692600
	5	21.55	1375	529400
H:1	1	14.76	1798.7	626115.4
	2	12.90	1950.1	682121.8
	3	12.82	1804.2	680025.8
	4	16.33	1753.8	624877.4
	5	20.23	1075	513130.4
H:2	1	17.02	1614.2	504199.4
	2	16.38	1764.6	604667.2
	3	14.45	1764.8	641493.4
	4	16.94	1750	598443.2
	5	19.98	1300	512235.4
H:3	1	2.81	1024	356231
	2	18.76	1316.9	509335.4
	3	16.44	1760.6	596402.6
	4	16.79	1750	579937
	5	19.86	1004.8	508073.6

in scenario 3 (0.36) than in scenario 4 (0.16). Since A2:G3 converts CD/DC plays into DD, in scenarios 3 and 4, a large percentage of the population achieves scores around 1000. There are therefore fewer intermediate scores between 1000 and 3000 and since the chance of establishing CC initially is higher in scenario 3 than scenario 4, this means that a small percentage of the population scores highly in scenario 3 whilst a large percentage score lower. Since CC becomes more implausible initially in the context of scenarios 4 and 5, more and more of the population score around 1000 resulting in the increased Gini coefficients observed in table 8.35

With reference to average median scores and average total scores an expected decreasing trend of both values for all non-greedy characters as scenario number is increased, is observed. This is due to the simple fact that the chance of establishing DD initially increases and the chance of establishing CC initially decreases as scenario number is increased. This ultimately results in lower individual and total system scores being achieved.

Greedy Emotional Characters The introduction of greed causes the results discussed for non-greedy emotional characters in relation to average Gini, median and total system scores to change. To help explain the results obtained, three graphs are provided for each emotional character that show the final scores of all agents in the first

Table 8.35: Average Gini coefficients, average median individual scores and average total system scores for all variants of A2:G3.

Hope Character	**Scenario**	**Avg. Gini Coefficient**	**Avg. Median Score**	**Avg. Total Score**
H:0	1	0.00	3000	1014000
	2	18.54	2599.2	771840.8
	3	24.42	1605.2	582164.8
	4	18.77	1008	447779.2
	5	7.51	1000	367840.8
H:1	1	1.20	1069	362031.8
	2	1.71	1044.1	353871.6
	3	1.46	1016	347343.4
	4	0.98	1008	342730.4
	5	0.50	1000	339810.4
H:2	1	0.24	1027.8	347490
	2	0.54	1019.4	344725.6
	3	0.54	1012	342328.6
	4	0.46	1008	340543.2
	5	0.29	1000	339022
H:3	1	0.11	1024	346161
	2	0.43	1020	343899.6
	3	0.47	1010	341888.8
	4	0.41	1008	340301.6
	5	0.28	1000	338981

repeat for each scenario. These graphs help to see why the Gini coefficients alter in the way that they do as the temptation to be greedy increases for each emotional character. The values in the graphs are sorted into ascending order and consequently, the values on the x-axes indicate arbitrary agent IDs. In considering the equality of greedy emotional character variants I will first discuss trends pertaining to each greedy character variant of an emotional character before discussing general trends between greedy variants of emotional characters.

A2:G1. If greedy variants of A2:G1 are first considered (see table 8.33 and figures 8.1, 8.2 and 8.3) then a decrease in the average equality of the majority of the populations considered is apparent as scenario number increases from 1 to 4. This is due to a smaller percentile achieving higher scores (as CC becomes less likely initially) and a larger percentage of the population achieving lower scores so the distribution of wealth throughout the population becomes more unequal. An increase in average equality is observed as scenario number increases from 4 to 5 since the likelihood of initially establishing DD outweighs all other potential initial plays. Since there are a huge number of agents establishing DD plays with their opponents initially this results in a concentration of scores around 1000, resulting in the observed increase in average equality (along with a significant reduction in total system score).

Analysing the effect upon average equality as the likelihood of greed increases is facilitated by calculating an average of the average Gini coefficients for each greedy variant of A2:G1 (see table 8.33). These meta-averages for each emotional character are shown below together with the average values from which they are calculated:

- A2:G1:H1 - Average of average Gini coefficients = 14.32 (using values 12.84, 13.32, 14.89, 16.13 and 14.41)
- A2:G1:H2 - Average of average Gini coefficients = 3.73 (using values 3.76, 3.87, 4.02, 3.95, 3.07)
- A2:G1:H3 - Average of average Gini coefficients = 5.02 (using values 2.98, 5.30, 5.87, 6.16, 4.80)

So, as the temptation to be greedy increases within the A2:G1 populations, it is observed that equality first increases and then decreases (this observation is also reflected in figures 8.1, 8.2 and 8.3[4]). Since the temptation to defect is at its lowest probability with A2:G1:H1, there are a wide range of scores obtained and therefore, the population experiences less equality. With respect to A2:G1:H2, all possible plays have an equal chance of occurring when hope is activated therefore, a large percentage of agents will establish CD/DC plays. Since A2:G1 converts such plays into CC, hope is guaranteed to be activated once again giving a chance for DD to occur. Therefore, more agents score around 1000 whilst scores obtained by other agents are lower. The decrease in equality observed as the temptation to be greedy increases is rather surprising since one would expect scores to become more equal because a greater percentage of the population achieves scores around 1000 (since the likelihood of establishing DD from CC is extremely high). However, this is not the case: as can be seen in figure 8.3, final scores for agents have a greater range than for A2:G1:H2 (see figure 8.2). Since CD/DC play is unlikely following the activation of hope, the scores of agents are ultimately decided by chance: agents either establish DD or CC when hope is activated. It may be that some agents are lucky in that CC is established when hope is activated, but this will only occur to a very small percentage of the population. A large majority of the population on the other hand will score around 1000. Consequently, there are few intermediate scores between 1000 and the highest score achieved and this explanation accounts for the results perceived.

With respect to average median and total system scores of greedy variants of A2:G1, A2:G1:H1 populations see a general decrease in both scores as scenario number is increased. For emotional characters A2:G1:H2 and A2:G1:H3 however, the average median individual and total score increases markedly from scenario 1 to 2 (hope is not guaranteed to be activated in round 4 due to ubiquitous initial cooperation in scenario 2 therefore some agents will achieve greater individual scores). Following scenario 2, 4b's average median individual and total score decreases as the scenario number increases to

[4]Larger versions of these tables can be found in appendix A, see tables A.1, A.2, A.3.

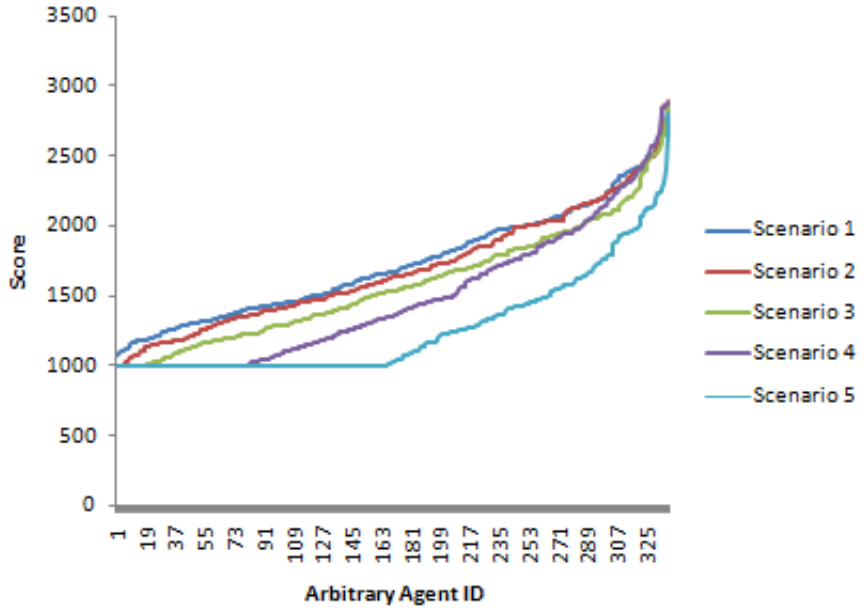

FIGURE 8.1: Final scores of A2:G1:H1 agents from the first repeats of scenarios 1-5.

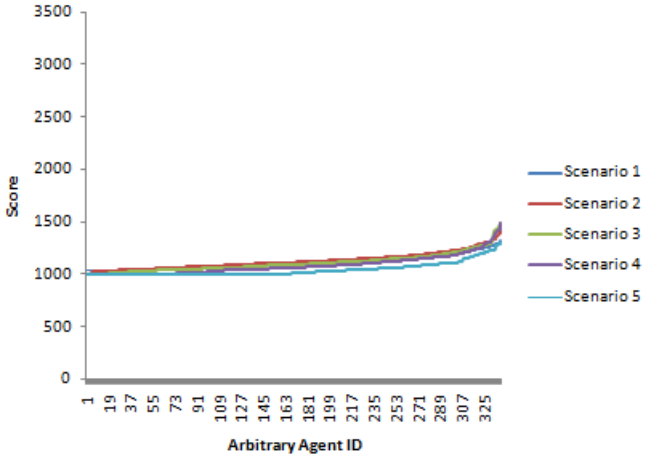

FIGURE 8.2: Final scores of A2:G1:H2 agents from the first repeats of scenarios 1-5.

scenario 5 as expected (as more and more agents lock into DD initially). For populations of A2:G1:H3 though, the situation is markedly different as the average median individual and total system score continues to increase from scenario 1 to 3 then decreases from 3 to 5.

This observed increase in average median and total system scores for A2:G1:H3 is relatively simple to explain: since all agents initially establish CC with each other in scenario 1, all agents are tempted by greed on round 4. Due to the likelihood of establishing DD after hope's activation is so high (0.81), most agents individual scores and the total score of the system, will suffer. As scenario number is increased from 1 to 2, CD/DC plays become more likely allowing some opponents to exploit others before establishing CC. This results in an increase with respect to both average total system

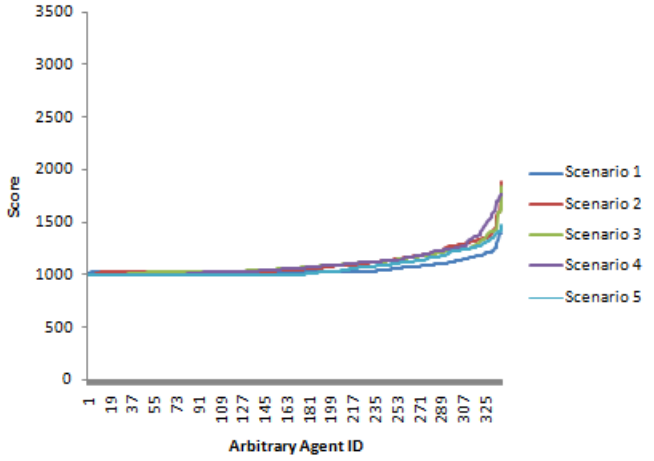

FIGURE 8.3: Final scores of A2:G1:H3 agents from the first repeats of scenarios 1-5.

and average individual scores because the total system can now earn up to 23 points from an agent and one of its opponents (DC followed by 3 CCs) compared to 18 (3 CCs, no CD/DC play). The more likely CD/DC is, the more this can occur (leading to an increase in individual score for more of the population). Such behaviour is not possible in the context of scenario 1 though since defection does not occur initially.

Yet, this does not account for why an increase in average individual score and average total system score occurs as scenario number is increased from 2 to 3 for A2:G1:H3 and not for A2:G1:H2. The answer to this lies in the fact that all plays are equally likely of being established by A2:G1:H2 resulting in more agents establishing CC and CD/DC plays in scenario 2. In this way, the average median and average total system scores of A2:G1:H2 are maximised in this scenario whereas for A2:G1:H3 they are reduced (due to the high likelihood of DD being established after hope's activation).

The decrease in average individual median and total system scores for A2:G1:H3 as scenario number increases from 3 to 5 occurs because the likelihood of establishing DD initially outweighs the chance of establishing CD/DC initially as is the case from scenario 4 onwards. As already explained, DD earns both the individuals and the total system the lowest scores possible and the more likely it is for agents to establish these plays initially, the greater the percentage of the population that scores lower.

A2:G2. With respect to the greedy variants of A2:G2, there is a general trend that holds where average equality increases from scenarios 1 to 3 but then decreases from scenario 3 to 5. The only exception to this rule is with A2:G2:H3 whose population is more equal on average in scenario 1 than in scenario 2. The explanation for this is due to the fact that ubiquitous cooperation guarantees the activation of hope and because the temptation to be greedy is so great many agents achieve scores around 1000. Furthermore, those agents who establish CD/DC plays after hope's activation maintain these plays; they are not capable of establishing CC and increasing their scores by any great amount. Consequently, the gap between the lowest and highest scores is relatively

small for A2:G2 populations compared to A2:G1 populations who can establish CC from CD/DC plays.

The observed trend that occurs for each greedy variant of A2:G2 is also explained by A2:G2's maintenance of CD/DC plays once established. Tables 8.4, 8.5 and 8.6 illustrate the dispersion of individual scores for A2:G2 populations with different likelihoods of greed[5]. As CD/DC play becomes more likely initially, a greater percentage of the population achieve similar scores as scenario number increases to scenario 3. The probability of CD/DC play occurring in scenarios 3 and 4 is equal however so the question must be asked: why is scenario 4 more unequal on average than scenario 3? By considering the likelihood of CC/DD occurring initially in these rounds, an answer can be provided. An agent earns 3 points each round when CC is established between itself and its opponent whilst only 1 point is earned by an agent when DD is established. The agent who defects in a CD/DC play however, earns 5 points per round so, there is less difference between individual points earned if CD/DC plays and CC are compared than there is if CD/DC and DD are compared. Since scenario 3 has a higher likelihood of CC being established initially then scores will be more concentrated than they would be if DD plays have a higher likelihood of being established initially. Essentially, the point to be made here is that the fairness of a system is increased if the majority play TFT and it is more likely that CC is established initially.

If the trends between greedy A2:G2 variants are now compared by calculating the meta-averages of the average Gini coefficients listed in table 8.34 then the effect of A2:G2's CD/DC maintenance upon population equality can be made clear:

- A2:G2:H1 - Average of average Gini coefficients = 15.91 (using values 14.76, 12.90, 12.82, 16.33, 20.23)
- A2:G2:H2 - Average of average Gini coefficients = 16.95 (using values 17.02, 16.38, 14.45, 16.94, 19.98)
- A2:G2:H3 - Average of average Gini coefficients = 14.93 (using values 2.81, 18.76, 16.44, 16.79, 19.86)

Increasing the likelihood of greed occurring from "less" to "moderately" decreases the equality of A2:G2 populations but increasing the likelihood of greed occurring from "moderately" to "highly" increases equality. Since A2:G2:H1 populations are less likely to establish DD or CD/DC plays from greed but are more likely to establish CC, the individual scores of agents within this population are fairly concentrated. However, A2:G2:H2 populations are equally likely to establish all kinds of plays and so, some agents may establish DD following greed, some may establish CD/DC and maintain this play and some may establish CC which may lead to CD/DC establishment and maintenance. This results in the potential for a wide variety of scores to be obtained for the A2:G2:H2 population, much more so than the variety of scores that may be

[5]Larger versions of these tables can be found in appendix A, see tables A.4, A.5, A.6.

potentially obtained in a A2:G2:H3 population. Under these conditions many agents will either establish DD initially or following greed whilst a very small number will establish and maintain CD/DC plays. Thus, the individual scores of agents within A2:G2:H3 populations are again concentrated, increasing equality.

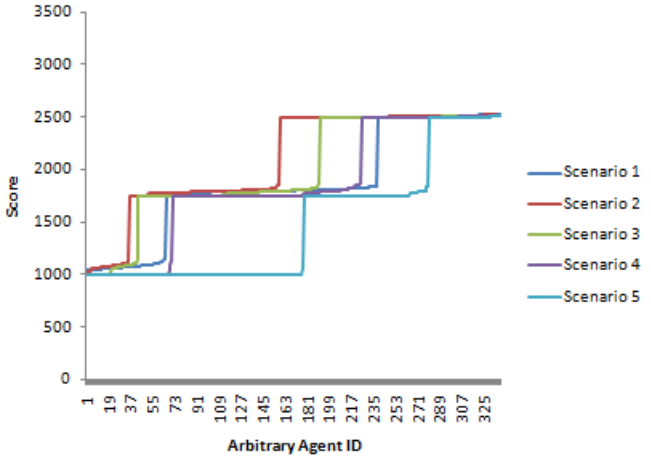

FIGURE 8.4: Final scores of A2:G2:H1 agents from the first repeats of scenarios 1-5.

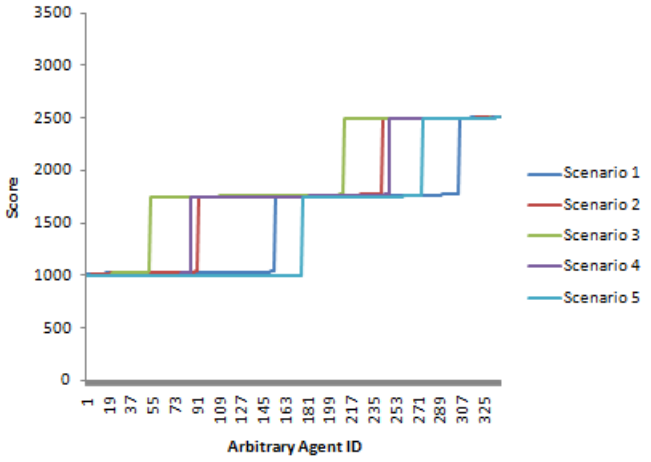

FIGURE 8.5: Final scores of A2:G2:H2 agents from the first repeats of scenarios 1-5.

Greedy variants of A2:G2 do not follow the trends in average median individual and total system scores as observed in its non-greedy emotional character variant. Average individual median and total scores increase for all greedy variants of A2:G2 as scenario number is increased from 1 to 2. This is for the same reason outlined in the discussion of average individual median and total system scores for greedy variants of A2:G1: hope is guaranteed to be activated for all agents in round 4 due to ubiquitous initial cooperation. Consequently, some agents will establish DD whilst some will establish CC resulting in hope being activated again (giving a chance of DD being established once again). Hope is not activated for all agents in scenario 2 however, since some agents may establish

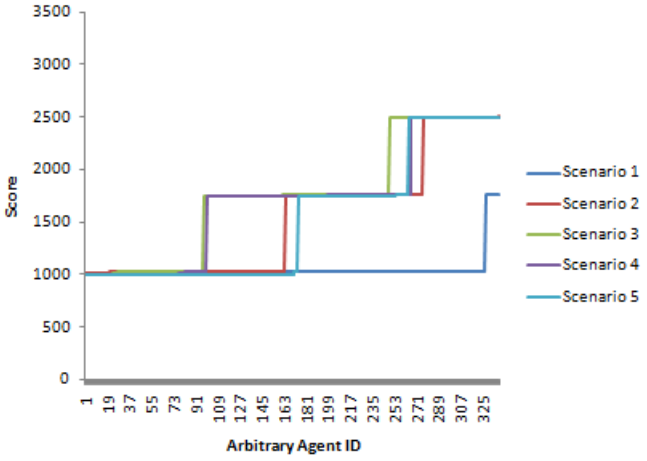

FIGURE 8.6: Final scores of A2:G2:H3 agents from the first repeats of scenarios 1-5.

CD/DC plays initially and A2:G2 will then maintain such plays earning individuals and the total system more points.

A2:G2:H2 and A2:G2:H3 experience a further increase with respect to average individual median and total system scores up until scenario 3, as A2:G1:H3 does (this does not occur with respect to A2:G2:H1). Again, this result can be explained by a consideration of initial play chances and play chances after greed is activated. In scenarios 3 and 4, CD/DC plays have more of a chance of occurring initially than CC/DD therefore, due to A2:G2's ability to maintain such plays, the average median individual and total score is increased. Compare this to scenario 2 where a large proportion of agents will establish CC initially resulting in the establishment of DD (0.5 chance for A2:G2:H2 and 0.9 for A2:G2:H3) and the explanation for the observation becomes clear.

The reduction in average individual median score and average total system score for A2:G2:H2 and A2:G2:H3 as scenario number is increased from 3 to 5 is explained by the fact that DD is more likely to be established than any other plays initially causing more of the population to achieve a score of 1000.

A2:G3. Generally speaking, for each greedy variant of A2:G3, there appears to be a trend where equality increases as scenario number increases from 2 to 5. This result is expected, along with the decrease in equality observed for each greedy variant of A2:G3 as scenario number is increased from 1 to 2. This decrease in equality between scenario 1 and 2 occurs because all agents are guaranteed to have hope activated in scenario 1 due to ubiquitous initial cooperation (resulting in many different plays being established). The difference between A2:G2 and A2:G3 however, is that agents convert CD/DC plays into DDs meaning that those who avoid the establishment of DD following hope's activation are all but guaranteed to establish DD if CD/DC plays are established instead. This results in a more equal system compared to A2:G2 but the observed increase between scenario 1 and 2 still exists for all greedy population variants. The establishment of DD from CD/DC plays also accounts for the trend of increased equality as scenario number is increased: initial DD becomes more likely along with CD/DC plays (up until scenario

4). In this case, all roads lead to DD meaning that, even though all agents may obtain low individual scores, the system itself is more equal.

Again, if the meta-average of average Gini coefficients is calculated to enable trends to be identified across greedy variants of A2:G3 then the populations become more equal as greed is increased. This is again due to emotional A2:G3's ability to convert CD/DC plays into DD meaning that all agents score close to 1000 (see figures 8.7, 8.8 and 8.9[6]). The meta-average Gini coefficients are shown below for reference but the point to be made here is that, if DD cannot be destabilised, then a system is more equal if individuals establish DD more quickly.

- A2:G3:H1 - Average of average Gini coefficients = 1.17 (using values 1.20, 1.71, 1.46, 0.98, 0.50)
- A2:G3:H2 - Average of average Gini coefficients = 0.41 (using values 0.24, 0.54, 0.54, 0.46, 0.29)
- A2:G3:H3 - Average of average Gini coefficients = 0.34 (using values 0.11, 0.43, 0.47, 0.41, 0.28)

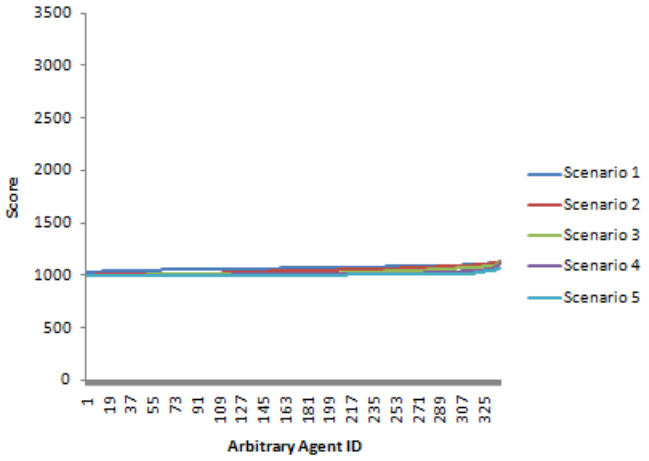

FIGURE 8.7: Final scores of A2:G3:H1 agents from the first repeats of scenarios 1-5.

With respect to average median individual and total scores only the greedy variants of A2:G3 follow the trend previously outlined for its non-greedy counterpart. This is because CC carries with it a chance of establishing DD and CD/DC plays are converted into DD by A2:G3. As the initial chance of CD/DC plays occurring increases as scenario number is increased from 1 to 3/4, as well as the likelihood of DD, the odds are heavily stacked in favour of DD occurring.

Social Orders As a final comment I intend to classify the emotional character populations considered in terms of social orders. Populations of A2:G3 agents appear to be

[6]Larger versions of these tables can be found in appendix A, see tables A.7, A.8, A.9.

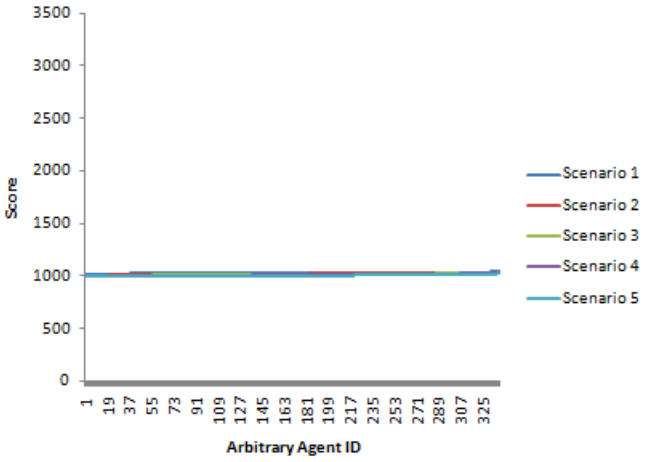

FIGURE 8.8: Final scores of A2:G3:H2 agents from the first repeats of scenarios 1-5.

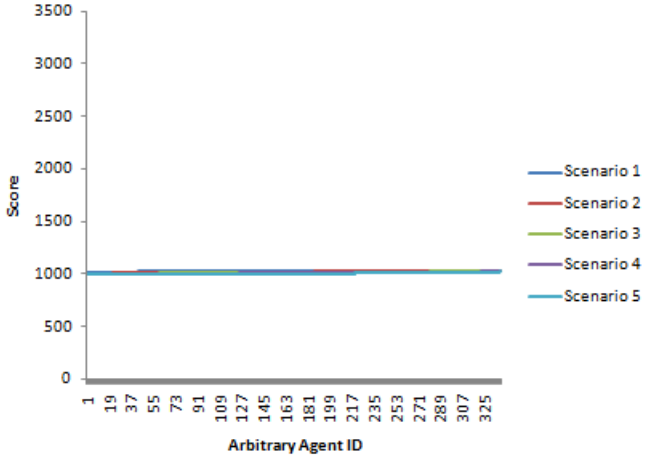

FIGURE 8.9: Final scores of A2:G3:H3 agents from the first repeats of scenarios 1-5.

akin to *command economies* due to their propensity to generate the lowest aggregate wealth whilst achieving the highest equality ratings. Populations of A2:G2 agents however appear to be typical *market economies* since aggregate wealth is high whilst equality is very low. Populations of A2:G1 agents could be described as a *mixed economy* where the freedom of the market to generate wealth is restricted through social safeguards (in this case greed). Another important point to note is that, in situations where DD can never be destabilised and CC carries with it a risk of greed, increasing the likelihood of greed occurring increases fairness. Likewise, the more likely greed is, the more players should sway towards establishing DD as this also increases the fairness of the system. As the cliché states: "*if you can't beat them, join them*".

8.6 Chapter Summary

In this chapter my intention was to create a computational mechanism inspired by human emotion to be implemented in a MAS simulation where agents playing an iterated

version of the Prisoner's Dilemma game could destabilise CC. *Hope* was identified as the human emotion to be used as a basis for the computational mechanism required and its effect is to enable *possible* periodic defection or *greed*. The simulations run investigated which of the nine basic emotional characters first described in chapter 6 are the most successful (both from an individual and total system standpoint) and why when initial defection rates are altered and the likelihood of hope's effect (greed) is increased/decreased. To ascertain the success of emotional characters given these varying conditions, three criteria were identified and considered, these are: *maximum individual score*, *minimum individual score* and *total system score*. The main findings from this investigation are:

- Ratios of tolerance and anger are extremely important and have a significant bearing upon the success of emotional characters in context of the scenarios run.
  - Emotional characters who are quicker to gratitude than anger (A2:G1, A3:G1 and A3:G2) establish CC from CD/DC plays.
  - Emotional characters who are equally quick to anger and gratitude (A1:G1, A2:G2 and A3:G3) maintain CD/DC plays if they are established.
  - Emotional characters who are quicker to anger than gratitude (A1:G2, A1:G3 and A2:G3) establish DD from CD/DC plays.
- In context of maximum and minimum individual scores:
  - When greed is non-existent or has a low likelihood of occurrence, A2:G1, A3:G1 and A3:G2 are successful with regards to maximising individual scores since establishing CC does not carry much of a risk of establishing DD through hope (greed). A1:G1, A2:G2 and A3:G3 are less successful than A2:G1, A3:G1 and A3:G2 but more successful than A1:G2, A1:G3 and A2:G3 because A1:G2, A1:G3 and A2:G3 have the highest likelihood of establishing DD following hope's activation.
  - When greed is moderately or highly likely to occur after hope's activation, A1:G1, A2:G2 and A3:G3 become more successful than A2:G1, A3:G1 and A3:G2. This is due to CC now carrying with it a significant risk of establishing DD after hopes' activation. Since A1:G1, A2:G2 and A3:G3 maintain CD/DC plays they earn more from an individual perspective and run no risk of establishing DD due to the activation of hope. A1:G2, A1:G3 and A2:G3 still perform poorly.
  - Increasing the number of initial defectors in the population appears to have no effect upon emotional character success.
- In context of total system scores:

  - The patterns of emotional character success observed for maximum and minimum individual scores are again observed for both non-greedy and greedy character populations and for the same reasons.
  - There is a deviation in the results observed when considering highly greedy emotional character populations in context of scenario 1. In this case, emotional character success is based upon *tolerance* rather than tolerance *and* responsiveness, otherwise DD becomes widespread.

Since the tolerance and responsiveness ratios of emotional characters appear to be fundamental to their success when periodic defection is possible in the simulations a further research question was posed i.e. "are there any behavioural features of emotional character populations with different tolerance and responsiveness ratios with respect to the *number of distinct scores obtained*, *percentages of the population obtaining these distinct scores* and *system equality*"? These questions were investigated using A2:G1, A2:G2 and A2:G3 (all equally tolerant but with different tolerance and responsiveness ratios) and the main findings are listed below:

- In context of distinct individual scores:
  - For non-greedy emotional character populations, A2:G1 and A2:G3 generate more individual distinct scores than A2:G2 since agents in these populations are capable of exploiting others before locking into CC/DD.
  - For greedy emotional character populations, A2:G1 generates more distinct individual scores than A2:G2 and A2:G3 whilst A2:G2 generates more distinct individual scores than A2:G3.
  - As the likelihood of agents being greedy is increased the number of distinct scores obtained by each emotional character population is reduced and the standard deviation of these distinct scores increases.
- In context of population percentages achieving these distinct scores:
  - The percentages of A2:G1 and A2:G3 populations scoring >3000 and <1000 are mirror images of each other.
  - As greed is introduced into populations, no agent is capable of achieving individual scores >3000 or equal to 3000 since CC cannot be established and maintained after exploitation or otherwise for an entire game. Therefore, greater percentages of A2:G1 and A2:G3 populations score <=2999 and >=1001.
  - A2:G2 populations have the greatest percentage of agents scoring <=2999 and >=1001.
- In context of equality:

- The non-greedy A2:G1 population is the most equal when the percentage of cooperators in the initial population is greater than the percentage of initial defectors. As more defectors are introduced, A2:G3 populations become most equal. A2:G2 populations are always most unequal.
- The introduction of greed into emotional character populations generally exaggerates these trends.
- A2:G2 populations tend to generate greater total wealth on average whereas A2:G3 populations generate less total wealth. A2:G1 produces total wealth in accordance with the likelihood of greed present within the population.
- If the emotional character populations considered were classified as social orders or markets then A2:G1 populations may be classed as *mixed economies*, A2:G2 populations may be classed as *market economies* and A2:G3 populations may be classed as *command economies*.

Chapter 9

Conclusions and Future Work

In a very broad sense, the aim of this thesis is to identify the importance of emotion in human decision-making in the context of public goods games and to propose how to model the eliciting conditions and effects of emotion in a computational manner. In doing so my goal was to study three main questions:

- What emotions should be modelled and how in a MAS context so as to enable the intentional behaviour of agents to be driven in a way that replicates human behaviour in public goods games?
- How should these emotions be modelled to enable the behaviour described above?
- How do these emotions affect societal interactions in the context of public goods games?

In providing answers to these questions many other issues required investigation. Firstly, consideration of human behaviour in public goods games necessitated a discussion of how standard game-theoretic interpretations of public goods games fail to explain the results obtained when these games are played by humans. Rather than the selfish, self-defeating behaviour predicted by game theory, human behaviour tends more to work in reverse to this prediction, especially in the context of public goods games and even if there is a cost to themselves for doing so. The cause of these results was then investigated and I proposed that *emotions* could be used to explain the behaviour observed in humans when playing such games. This discussion is presented in its entirety in chapter 2.

Following on from this in chapter 3, philosophical issues concerning emotion were considered with the question of whether a computer can experience an emotion proper and when emotion occurs in respect to human decision-making. If it is to be posited that emotion drives the behaviour of humans in public goods games then it must first be asserted that emotion occurs *before* behaviour. A discussion and evaluation of evidence that argues for emotion occurring before or after behaviour discovers that there are plausible arguments for emotion occurring in each of these temporal dimensions

with respect to behaviour. The relationship between emotion and *rationality* was then considered as, to many people, the two concepts oppose each other. With regards to computational agents, this thesis asserts that, if an agent is said to be rational, then the agent must behave in a way consistent with some knowledge base. Therefore, if a person performs an action that confers some benefit upon me, and I help them at some cost to myself afterwards, then my behaviour can be described as "rational" because I am *grateful* to that person due to the action they performed. I can then be said to be acting both *emotionally* and *rationally* (emotionally since my behaviour is driven by the emotion of gratitude and rationally because helping another at some cost to yourself is usual behaviour if somebody does the same for you). Given this assertion, if the concepts of emotion and rationality are graphically plotted, they are not diametrically opposed ideas. Instead, emotion and rationality create a two-dimensional (rather than one-dimensional) space where any point on this graph can be described in context of both emotion *and* rationality. Including emotion into computational systems therefore does not mean that such computational systems are irrational, it simply means that they are capable of both rational and emotional behaviour.

This philosophical consideration of emotion was then furthered in chapter 4 by an investigation of how emotion has been modelled in computational systems. Initially, major psychological models of emotion were discussed (appraisal, dimensional and anatomical) and it was ultimately decided that using an *appraisal* model of emotion would be the most appropriate for my purposes. Of the numerous appraisal models discussed, the *Ortony, Clore and Collins* (OCC) model was selected as the psychological emotional model to be used to develop computational emotions for the agents in this thesis. This decision was made due to the OCC's close correlation with agent-theoretic notions and its central assumption that it is the *agency* of a system that enables emotional episodes to be elicited. In the context of computer science, it emerged from a consideration of literature regarding the implementation of emotion in computational systems that emotions are usually modelled to be *functional* i.e. there is some input to the emotion's activation and activation of the emotion entails some discernible effect upon the decision-making and action-selection of the agent. A review of this literature also led to the discovery that *simulation* is the most frequently used method of investigating the effects of emotion in computational systems. Furthermore, there has been little research related to the investigation of how emotions affect societal interactions in MAS contexts. It was then decided that this was the general research question to be pursued.

A framework for modelling emotion in agent systems was proposed in chapter 5 along with the fundamental constituents of the simulations to be used and the general research agenda. With respect to modelling emotions it was proposed in chapter 5 that emotions can be implemented by specifying four *quantitative* and two *qualitative* elements. These elements, their type (qualitative/quantitative), a description of the element and an example of the element in the context of this thesis is provided in table 9.1.

TABLE 9.1: The six elements identified in this thesis required for modelling an emotion along with their type (qualitative/quantitative), a description of the element and an example of the element in the context of this thesis.

Element	**Type**	**Description**	**Example**
Eliciting condition	Qualitative	An event that causes the potential of an emotion to increase/decrease.	Receiving defection from an agent causes an agent's anger potential to increase by 1.
Potential	Quantitative	A numerical value that indicates how excited an emotion currently is.	If an agent, y, has cooperated with an emotional agent, x, x's gratitude potential may be set to 1.
Activation Threshold	Quantitative	A numerical value that, once equalled or exceeded by the emotion's potential, causes the effect of an emotion to become manifest in the agent.	If an emotional agent, x, has a gratitude activation threshold of 1 then, if x's gratitude instance potential towards an opponent y is 1, gratitude is activated in x towards y.
Saturation	Quantitative	A numerical value that indicates the maximum value that an emotion's potential can be increased to. No negative saturation value is specified since when an emotion's potential equals 0, its potential cannot decrease any further.	If an emotional agent, x, has a saturation value of 3 for gratitude then, if x receives more than three consecutive cooperations from an opponent y, x's gratitude instance potential towards y will not increase to 4, it will stop at 3.
Effect	Qualitative	An action that is mapped to the activation of an emotion.	If an emotional agent, x, has its gratitude instance potential towards y activated (due to its potential = activation threshold), x will cooperate with y in the next round.
Probability of effect	Quantitative	How likely is it for the effect of an emotion to be manifest in an agent (ranges from 0 to 1; 1 indicates that when the emotion is activated the effect is guaranteed to be manifest).	If an emotional agent x, has a probability of effect for gratitude equal to 0.5 then whenever any instance of x's gratitude is activated, its effect only has a 0.5 chance to be manifest in each round.

Table 9.2: A summary of notable results from chapter 6.

Chapter	Notable Results
6	- Total system score of an initially cooperative A3:G1 agent surpasses total system scores achieved by all other emotional and non-emotional agents for many reasonable initial configurations. - A3:G1's success is due to high *responsiveness* and *tolerance*. - A3:G1's qualities are especially important when playing against agents whose behaviour is uncertain. - As responsiveness and tolerance increase in an agent, fairness of scores suffer. - If tolerance and responsiveness ratio leads to a score of less than 200, an agent may as well defect constantly since this results in a score of 200.

Attention was then focused upon developing emotions using this framework that could be used to enable agents to play an iterated Prisoner's Dilemma game in chapters 6, 7 and 8. This public goods game was chosen in particular due to the limited actions available to participating agents. Such a feature facilitates the precise identification of eliciting conditions and effects for the emotions to be modelled. Three investigations were then undertaken resulting in the modelling of four emotions along with an in-depth analysis of how these emotions affect the social interactions of agents playing the iterated Prisoner's Dilemma game.

The emotions of *anger* and *gratitude* were first modelled in chapter 6 as a basis for motivating an agent's decision to cooperate or defect. This lead to the identification and implementation of nine emotional *characters* that would form the basis of the remaining investigations. Variance in the activation thresholds of anger and gratitude produced different *tolerance* and *responsiveness* ratios with these ratios being especially important in interpreting the performance of agents within these simulations. This answer was reached by a consideration of whether agents endowed with these emotions offer any improvement upon the success of agents endowed with the TFT strategy when playing against other notable strategies from Axelrod's famous computer tournament. To answer this, I queried how anger and gratitude affect rates of cooperation/defection and the fairness of individual scores when their activation thresholds are altered. Table 9.2 shows the notable results extracted from this set of simulations succinctly.

In the next set of simulations discussed in chapter 7, *admiration* was modelled and used to augment the nine emotional characters introduced in chapter 6. An investigation into what emotional characteristics are selected for under different initial conditions and why, was performed. This examination provided a discussion regarding what emotional social norms emerge in a population when agents admire the individual success of others. Two broad questions were then asked: "*do different initial conditions affect the emotional characteristics selected for?*" and "*is it is the case that emotional characters that promote the total wealth of the system are selected for as an emergent property?*". Notable results obtained in this chapter are included in table 9.3.

TABLE 9.3: A summary of notable results from chapter 7.

Chapter	Notable Results
7	- Increasing rates of initial cooperation in a population places a premium upon lesser tolerance and moderate responsiveness. - Increasing rates of initial defection in a population places a premium upon increased tolerance and responsiveness. - Increasing the initial percentage of less impressionable agents in a population causes highly tolerant and highly responsive emotional characters to become more prevalent. - Increasing the initial percentage of highly impressionable agents in a population causes no discernible effects to be observed. - Increasing the number of agents in any agents player and comparator sets exaggerates and concentrates emotional character prevalence; increasing the number of agents in the player set achieves this more so. - Highly tolerant emotional characters are admired more and tolerance is more admired than responsiveness. - A3:G1 is the most prevalent emotional character overall and its total score maximisation is an *emergent* property.

Finally, in chapter 8, I modelled *hope* so that agents were capable of periodic defection after CC had been established (a form of greed). Hope again augments the nine basic character types identified in chapter 6 and the simulations constructed allowed me to answer the question of how homogeneous emotional character populations handle destabilisation of CC cycles after their establishment. Three groups of distinct tolerance and responsiveness ratios were identified from a consideration of the results obtained: emotional characters that are more responsive than tolerant (A2:G1, A3:G1 and A3:G2); emotional characters that are equally responsive and tolerant (A1:G1, A2:G2 and A3:G3); and emotional characters that are less responsive than tolerant (A1:G2, A1:G3 and A2:G3). The dividing feature of these emotional character groupings is how they handle CD/DC plays:

- A2:G1, A3:G1 and A3:G2 establish CC once CD/DC plays are established.
- A1:G1, A2:G2 and A3:G3 maintain CD/DC plays once established.
- A1:G2, A1:G3 and A2:G3 establish DD once CD/DC plays are established.

Following this division of emotional character groups I ran a number of simulations to ascertain which of the nine basic emotional characters first described in chapter 6 are the most successful (both from an individual and total system standpoint) when initial defection rates are altered and the likelihood of hope's effect (greed) is increased/decreased and why. To answer these questions three criteria were identified and considered, these were: maximum individual score, minimum individual score and total system score. Since the tolerance and responsiveness ratios of emotional characters appear to be fundamental to their success when periodic defection is possible in the simulations, a further

research question was posed i.e. "*are there any behavioural features of emotional character populations with different tolerance and responsiveness ratios with respect to the number of unique scores obtained, percentages of the population obtaining these unique scores and system equality?*". These questions were again investigated by using a number of simulations and only considering emotional characters A2:G1, A2:G2 and A2:G3 (all equally tolerant but A2:G1 is highly responsive, A2:G2 is moderately responsive and A2:G3 is less responsive). Notable results from this chapter are listed in table 9.4.

By considering the above conclusions I believe that the rather broad research question outlined in section 1.1 of chapter 1 i.e. *"can emotions be modelled functionally in a computational context so as to understand their impact upon the social interactions of agents engaged in public goods games?"* has been answered. I have provided a general framework for modelling emotions in a functional context and an extensive discussion on how the emotions of *anger*, *gratitude*, *admiration* and *hope* affect social interactions within the context of a public goods game namely, the Prisoner's Dilemma. There is obviously much scope for the analysis of how other emotions affect social interactions in context of the Prisoner's Dilemma and in other contexts and this is discussed more in section 9.2 below.

The remainder of this final chapter is divided into three sections: section 9.1 outlines the major contributions and how they were achieved, section 9.2 discusses the limitations of the work performed in this thesis and suggests future work based upon these limitations that may be performed and section 9.3 offers some final words upon the applicability of this work to current world events and a summary of the work performed.

9.1 Major Contributions

Throughout this thesis my aim has been to provide novel and interesting contributions to a number of fields including philosophy, MAS, game theory and affective computing. Consequently, there have been several contributions made to these fields and some interesting and novel discoveries made as a result of the modelling and simulation of emotions. In outlining both the major contributions and how they were achieved I hope to clarify the novelty of this thesis and its context with respect to the advancement of the current state of the art.

- *I have shown that, rather than playing an adjunct role in decision-making, emotions are viable mechanisms to aid decision-making and implementation of intentional behaviour by taking into account future consequences of current actions. Therefore, incorporating emotions into agent-centric decision-making can help to prevent the self-defeating outcomes proposed by game theoretic studies of public goods games.*

This contribution is composed of a number of arguments proposed in chapters 2 and 3. In chapter 2, I argued for emotions being an integral part of human decision-making

TABLE 9.4: A summary of notable results from chapter 8.

Chapter	Notable Results
8	- When greed is non-existent or has a low likelihood of occurrence, A2:G1, A3:G1 and A3:G2 are most successful at maximising individual/total scores; A1:G1, A2:G2 and A3:G3 are less successful; A1:G2, A1:G3, A2:G3 are least successful. - When greed is moderately or highly likely to occur, A1:G1, A2:G2 and A3:G3 are most successful at maximising individual/total scores; A2:G1, A3:G1 and A3:G2 are less successful; A1:G2, A1:G3 and A2:G3 are least successful. - Increasing the number of initial defectors in the population appears to have no effect upon the success of emotional characters at maximising individual scores. - In scenario 1 when greed is highly likely to occur, emotional character success is based solely upon *tolerance* rather than tolerance *and* responsiveness. - For non-greedy emotional character populations, A2:G1 and A2:G3 generate more individual unique scores than A2:G2. - For greedy emotional character populations, A2:G1 generates more unique individual scores than A2:G2 and A2:G3; A2:G2 generates more unique individual scores than A2:G3. - Increasing the likelihood of agents being greedy reduces the number of unique scores obtained by each emotional character population and increases the standard deviation of these unique scores. - The percentage of A2:G1 and A2:G3 populations scoring >3000 and <1000 are mirror images of each other. - After introducing greed, no agent achieves individual scores >3000 or equal to 3000 but greater percentages of A2:G1 and A2:G3 populations score <=2999 and >=1001. - A2:G2 populations have the greatest percentage of agents scoring <=2999 and >=1001. - Non-greedy A2:G1 population obtains the most equal individual score distributions when there are more initial cooperators than initial defectors; as initial defection increases, A2:G3 obtains the more equal individual score distributions; A2:G2 populations always obtain the most unequal individual score distributions; increasing greed likelihood generally exaggerates these trends. - A2:G2 populations generate most total wealth on average; A2:G3 populations generate least total wealth on average; A2:G1 populations generate total wealth in accordance with the likelihood of greed. - A2:G1 populations may be classed as *mixed economies*; A2:G2 populations may be classed as *market economies*; A2:G3 populations may be classed as *command economies*.

by presenting experimental research from a number of different disciplines. This research illustrates how human players in public goods games are more likely to *cooperate* with one another than defect, especially when some form of social interaction has occurred before the game is played. Such results oppose the predictions made by game theory whereby defection is the rational strategy for players to employ. The observation that social interaction exerts such an effect leads into a presentation of research that argues for emotion being the primary motivation for why human players do not act in accordance with game theory predictions. The core ideas presented identify emotions as being guarantors as credible threats and rewards (which affects subsequent actions of the opponent) and posits that the actions of the human players studied are motivated by emotion.

Chapter 3 then presents three major philosophical arguments that underpin the notion that emotions can and should be modelled in agent systems to facilitate agent-centric decision-making and intentional behaviour that prevents the self-defeating outcomes of game theory occurring in public goods games such as the Prisoner's Dilemma:

1. Emotions occur *prior* to intentional behaviour.
2. Emotions are capable of *motivating* intentional behaviour.
3. Emotions *do not always* give rise to intentional behaviour that is irrational.

These claims are buttressed by major philosophical, psychological and neurological research that argues for their validity. Based upon these arguments it can therefore be asserted that modelling agent-centric decision-making and intentional behaviour upon observations of the causes and effects of human emotion in public goods games is a viable idea from a computer science perspective. This is due to the fact that the research considered in chapters 2 and 3 gives evidence to assert that reasonable behaviour, which makes sense in human terms, can be driven by emotions with no reference to pay-offs, play histories or predetermined strategies.

- *I have developed a test-bed that may be used to investigate the effects of emotions upon societal interactions in context of an iterated Prisoner's Dilemma game given differing initial conditions.*

Following consideration of a number of major pieces of computer-science research concerned with computational implementation of emotions, it was discovered that little research exists with respect to the effects of emotion upon societal interactions in MAS games (see chapter 4). Therefore, to perform research into this topic, the methods used to investigate similar questions was taken into account in chapter 4. In the majority of cases the works in question use *simulations* to test an emotional model's ability to either:

- Provide immediate responses for computer systems that are situated in competitive, resource-bounded environments.

- Mirror experimental observations of emotions in human beings.
- Allow the computer systems that the emotional model is implemented in to display emotion so as to suspend belief in human-beings that such computer systems are machines.

Simulation is well suited to such investigations as they allow for the precise control of independent variables, ease of experiment repeatability and efficient collection of data. Consequently, the decision was taken to develop a simulation test-bed to answer the main research questions listed in the introduction to this chapter. The general features of this simulation test-bed developed are outlined in chapter 5 but from a general perspective, the test-bed is a multi-agent simulation where various initial conditions can be modified. By doing this, I attempted to facilitate the investigation of the effects emotions have upon societal interactions between agents in a MAS played an iterated version of the Prisoner's Dilemma game.

- *I have modelled various important emotions with respect to the iterated version of the Prisoner's Dilemma game.*

The emotional modelling framework proposed in this thesis is presented in chapter 5 where the following basic elements were defined as being essential to the computational modelling of any emotion.

- Eliciting condition(s)
- Activation threshold
- Effect
- Probability of effect

These elements were identified after a review of the literature concerning how emotions can be modelled from both a psychological and computational perspective. Note that the specifics of these elements need not be defined: the agent's own internal belief structures etc. are responsible for the specifics of the elements outlined. The elements identified for the emotional model were then used to model and endow agents with several emotions that drive intentional behaviour in human beings in the context of public goods games, specifically the iterated Prisoner's Dilemma. The details of how these emotions are modelled and why, are provided in chapter 6 (anger and gratitude), chapter 7 (admiration) and chapter 8 (hope).

- *I have provided a comprehensive study into the effects of anger, gratitude, admiration and hope upon societal interactions between small and large scale agent populations under differing initial conditions.*

To be exact, three comprehensive studies concerning the effects of anger, gratitude, admiration and hope upon societal interactions were run and reported upon in chapters 6, 7 and 8. In each of these simulations the effects of the emotions given different initial conditions such as the percentage of initial cooperators/defectors, percentages of highly/moderately/less tolerant, responsive, impressionable and greedy agents etc. were investigated. One should be mindful that the major conclusions given in these chapters have a specific context i.e. all results pertain to agents playing the iterated version of the Prisoner's Dilemma. Generalising these results to give some prediction of behaviour for agents capable of experiencing anger, gratitude, hope and greed in other public goods games is possible. However, the exact modelling of these emotions in such games has not yet been investigated so the behaviour predicted may not be exact. Therefore, caution should be taken if using these results to explain observations made in respect to emotional agents playing public goods games other than the iterated Prisoner's Dilemma.

9.2 Limitations and Future Work

The work contained in this thesis is extensible since the number of emotions capable of being computationally modelled and the contexts in which they can be studied can be extended since only four of the twenty-two OCC emotions [142] have been considered. In this section I will outline the limitations of the work performed and from these, propose possible extensions that would advance the current state of the art. These suggestions provide novel avenues of research for others who may be interested in the subject of emotional agent systems and the study of simulated social interactions using computational models of emotion.

Firstly, from a theoretical perspective, decision-making using emotional mechanisms should be more efficient and reduce non-determinism in highly unpredictable environments but this is not actually investigated in this thesis. Such research would be classed as *cognitive-engineering* research according to Burghouts et al. [25]. An investigation like this could be performed by benchmarking the performance (time taken to make a decision, correctness of decision etc.) of emotional computational mechanisms against the performance of non-emotional mechanisms in highly dynamic environments.

Following on from this suggestion, one could analyse the effects of emotion upon societal interactions in other MAS environments. In section 5.3 of chapter 5, I noted that the emotion model developed in this thesis was context-specific to the iterated Prisoner's Dilemma. By using different public goods games, an understanding of how emotions affect social interactions could be developed leading to improved effectiveness and applicability of emotion implementation in MAS. The limitation imposed by the particular simulation context could also be alleviated considerably by augmenting the existing emotional model so that agents are capable of appraising events and actions in a much more generic fashion (currently, the agent's emotional pre-conditions and

effects are essentially hard-coded). An interesting avenue of research would be to allow agents to develop their own goals, standards and attitudes and see how their emotional responses affect the societal interactions they engage in.

The effect of emotions upon prior and post decision-making is not addressed in this thesis. As argued for by Fehr and Gächter in [56], potential free-riders anticipate that their actions will cause anger and therefore punishment from others so will refrain from acting greedily for fear of punishment. This type of reasoning could realistically reduce the chance of periodic defection occurring by way of autonomous emotional reasoning since, whilst free-riding may evoke negative emotions and punishment from others, negative emotions may also be elicited in the free-rider further dissuading them from employing greedy behaviour. As postulated by Baumeister et al. in both [11] and [10], emotions may also "feedback" on decision-making to promote or demote certain intentional behaviours in particular situations in future. Again, modelling emotion in this way could also help to promote cooperation and reduce unfair pay-offs in the context of the Prisoner's Dilemma.

As demonstrated by Frank et al. in [64], social interactions between human players prior to public dilemma games appear to increase the rates of cooperation observed between experiment participants. Unfortunately, although this period of social interaction was not modelled in the simulations used in this thesis, the effects of such interactions are interesting. By allowing opponents to talk prior to playing public goods games, it would appear that emotional threats and promises are given further credence. Frank et al. observed that rates of cooperation increased in their investigations when participants were given the chance to meet their opponents before the games designed were played. By simulating such interactions important aspects of social encounters may be revealed aiding the development of emotional agents that independently choose to cooperate with others in competitive environments.

Related to the above proposal is the idea that selection of emotionally-motivated behaviours is not affected by considerations of how others may evaluate such behaviour. This influence is noted by Scherer in [167] and is mentioned in section 4.1.1.4 of chapter 4. By taking into consideration the evaluation of actions by others, even more truthful models of emotion may be developed. This would create interesting new dynamics with regards to agent decision-making and may even enable agents to engage in behaviour that may be construed as having some form of "political" agenda.

The emotional modelling framework and the emotions designed using it in this thesis are somewhat limited since there is no consideration given to intensity variables of emotions. This is a complex issue with many psychologists disagreeing about what variables affect the intensity of certain emotions and how. By modelling variations of intensity, social interactions between agents are capable of becoming richer and more true to life, potentially enabling researchers to accurately predict and explain the behaviour of individuals in reality by using agent-based simulation. For example, if an agent receives the sucker's pay-off then its anger potential could increase more than it would if both the

agent and its opponent defected mutually. The action is the same but the consequences are different, such considerations would make for especially interesting results.

Unlike anger and gratitude, admiration and hope were not modelled along with their counterparts: reproach and fear. Modelling and implementing these emotions in MAS and analysing their effects upon social interactions would again increase the richness of social interactions observed and further the understanding of how emotions affect these social interactions. Both emotions make for interesting research opportunities but with particular reference to reproach, this emotion would ensure that unsuccessful emotional characters have a greater selection pressure placed upon them, making more successful emotional characters become more prevalent (possibly increasing the amount of cooperation within a system as an emergent property).

A further limitation of this work is the absence of an emotional decay function. Whilst it was argued in section 5.3.4 of chapter 5 that emotional decay is not considered because emotion eliciting events occur so temporally close to one another in the Prisoner's Dilemma game that decay would not reasonably have any chance to exert any effect, it may be that constant activation of emotions (resulting from CC/DD) results in a decrease of that emotion's potential over time. Such an observation can be verified by anecdotal evidence: if you were to receive a similar gift every day from the same person, the event would eventually become normalised and the duration for which gratitude was felt when the first gift was received would decrease over time as the event occurs repeatedly. Also in situations where, for example, an A3:G2 agent receives two defections followed by two cooperations and then another defection, it may be that this third defection would not feasibly activate anger in the agent since the two intermediate cooperations would have allowed the agent's previous anger levels to decrease. Developing and implementing an objectively correct emotional decay function would allow for more realistic and rigorous simulations to be constructed which would, in turn, increase the validity of any results obtained by using them.

Emotional contagion (mentioned in section 5.3.3 of chapter 5) is also not considered in this work. Modelling such a phenomenon would again allow for more realistic and rigorous simulations and more valid results to be obtained. As mentioned in the reference above, attempts have been made to model emotional contagion by Masthoff and Gatt [124] and this would be a good starting place for anyone interested in incorporating such a feature in the emotional modelling framework used in this thesis.

In chapter 7 it was noted that 338 agents are used and this number is not equally divisible by 9 which results in an unequal number of emotional characters present in the initial population. Whilst this may not make much of a difference given that the difference between numbers is so small (1 or 2), it should rightly be listed as a limitation of the work. If these results were to be repeated with a total number of agents equally divisible by 9 it would be interesting to ascertain whether this makes much of a difference to the results obtained already.

Also in chapter 7, there is no comment made upon any interplay between initial defection percentages and percentages of admiration levels in the simulations. It may be that there are some interesting comments to be made with respect to this but the set-ups used did not facilitate such an investigation. This could be addressed in any future work by creating initial set-ups that vary both the initial percentage of defectors and admiration levels in the populations.

9.3 Final Words

Social interactions and the effects of emotion upon them is, in my opinion, one of the most pertinent and interesting avenues of emotional computing that still remains relatively unexplored; especially in the context of agent systems. The work presented in this thesis only begins to illuminate a small area of a vast topic. As can be observed from section 9.2 above, there are many paths open to those who are willing to follow them. It is my intention that the work I have undertaken and documented may act as an infrastructure for those who wish to perform further research in this field.

Currently, social dilemmas are rife: the recent exposés regarding multinational corporations not paying their "fair-share" of taxes [193], [22] is a classic example of a Prisoner's Dilemma game that occurs in reality. By not paying its "fair-share" of tax a corporation can be accused of free-riding or defecting, but why does perceived free-riding provoke such a strong reaction? The answer, as I have proposed throughout this work, is rooted in *emotion*. By understanding why certain emotions are elicited in certain situations and how these emotions affect intentional behaviour, a better understanding of the relationship between emotions and social interactions can emerge. These insights can serve to promote cooperation and reduce defection in social dilemma contexts but such insights must first be obtained. It is my hope that the major contributions of this work, listed in section 9.1, can be used to aid the acquisition of such insights and provide a starting point for those interested in the topic. As noted by Plato: "*human behaviour flows from three main sources: desire, emotion, and knowledge.*"; it is my hope that this work outlines the importance of emotion in determining human behaviour since rational, game-theoretic models appear to fall somewhat short of the task.

In this thesis I have identified: a methodology to investigate how emotions affect social interactions; a standardised framework for modelling emotion in agents; the basic emotions involved in human decision-making in a Prisoner's Dilemma game context; how emotion affects the establishment and maintenance of cooperation between agents in an iterated Prisoner's Dilemma game context; what emotional characteristics are selected for in a population of emotional agents playing the iterated Prisoner's Dilemma game; how emotion can affect the proliferation of cooperation in a population of agents playing the iterated Prisoner's Dilemma game; how different ratios of tolerance and responsiveness respond to periodic defection in an iterated Prisoner's Dilemma game; and what

effect (if any) various initial conditions have upon emotions and social interactions. I believe that the above work makes a significant and interesting contribution to the current state of the art. As stated earlier, there are still many questions to be answered before a complete understanding can be achieved of how emotion affects social interactions, though this thesis provides a solid grounding with results that can be taken forward.

Appendix A

Chapter 8 Charts

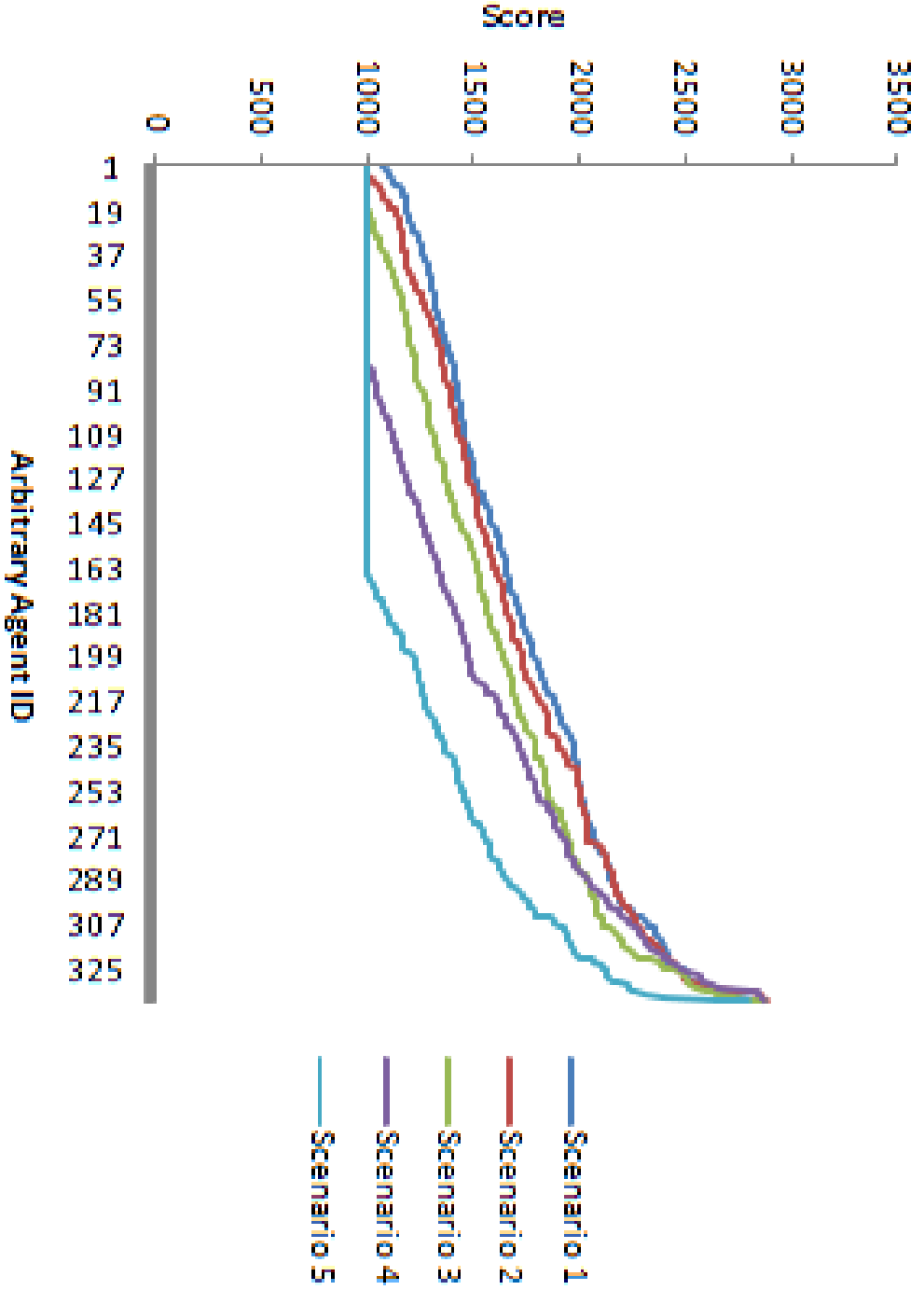

FIGURE A.1: Final scores of A2:G1:H1 agents from the first repeats of scenarios 1-5.

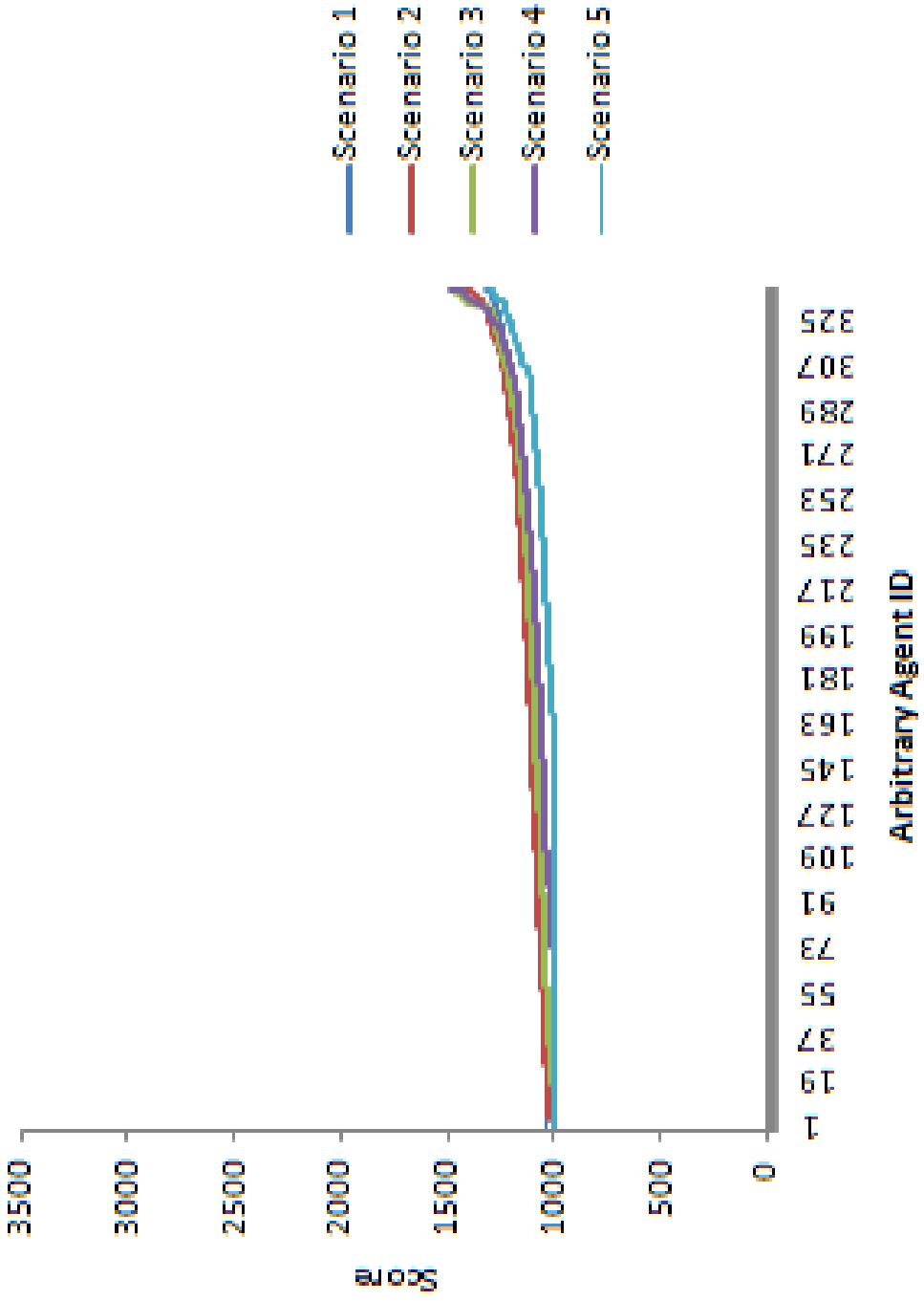

FIGURE A.2: Final scores of A2:G1:H2 agents from the first repeats of scenarios 1-5.

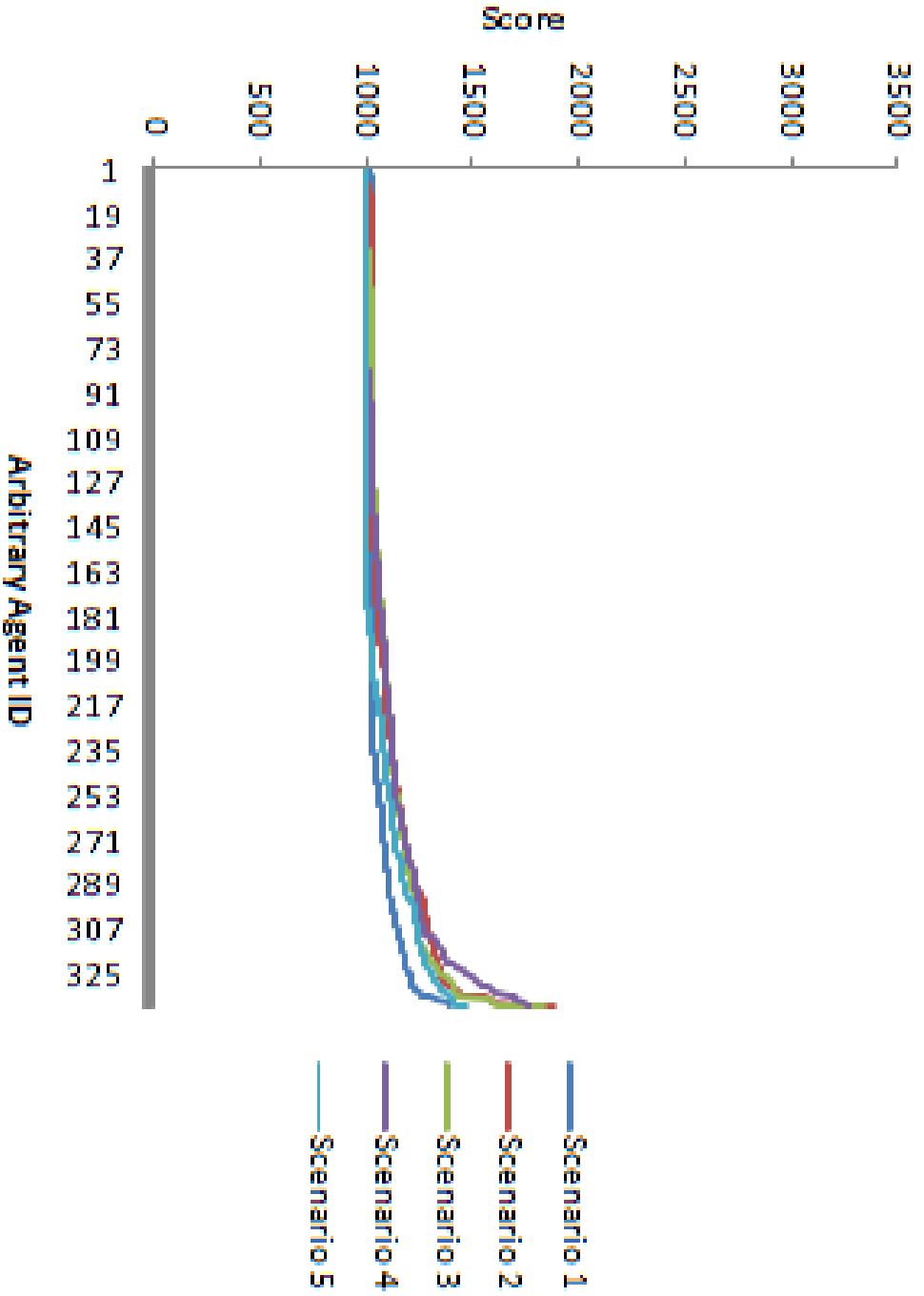

FIGURE A.3: Final scores of A2:G1:H3 agents from the first repeats of scenarios 1-5.

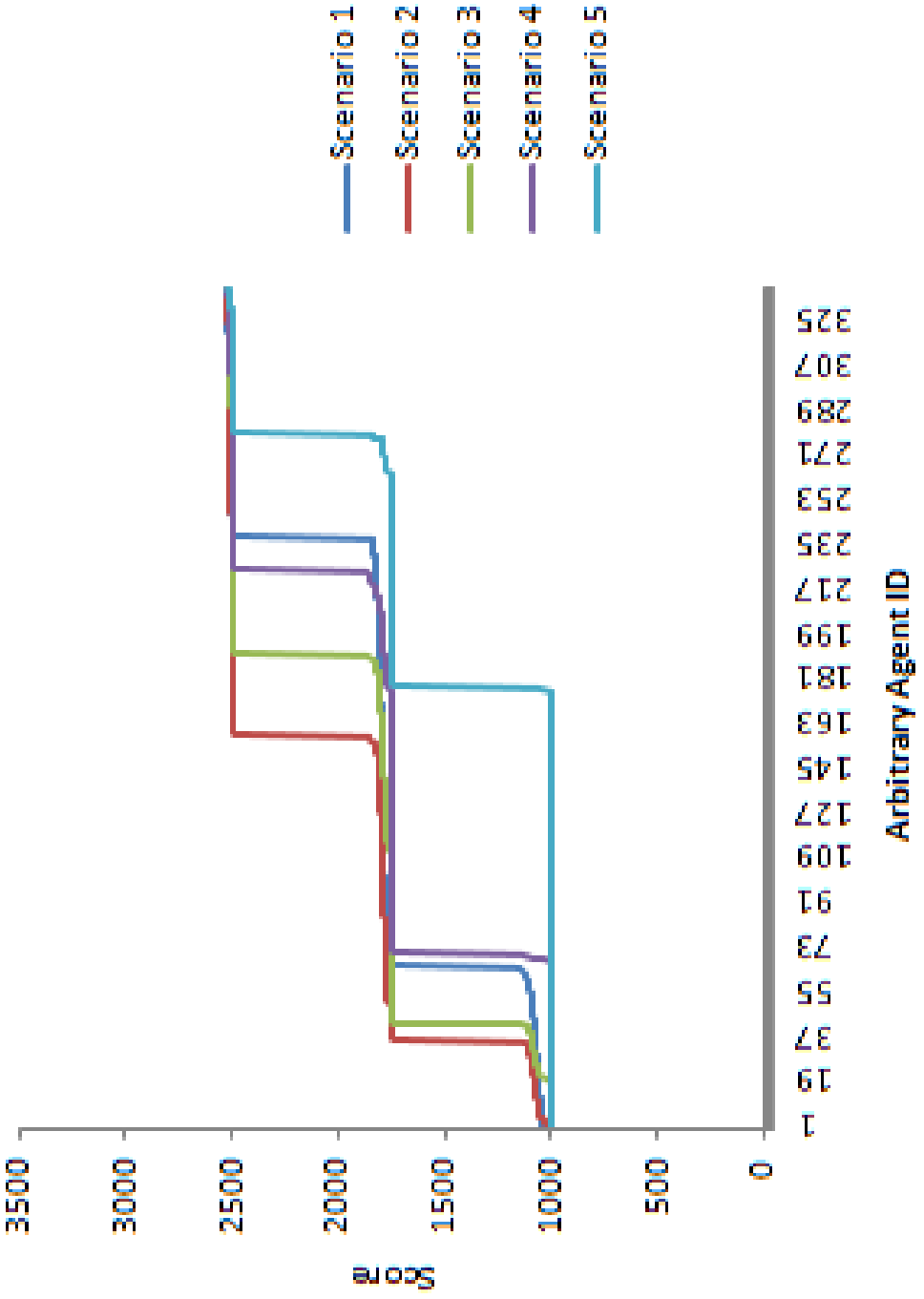

FIGURE A.4: Final scores of A2:G2:H1 agents from the first repeats of scenarios 1-5.

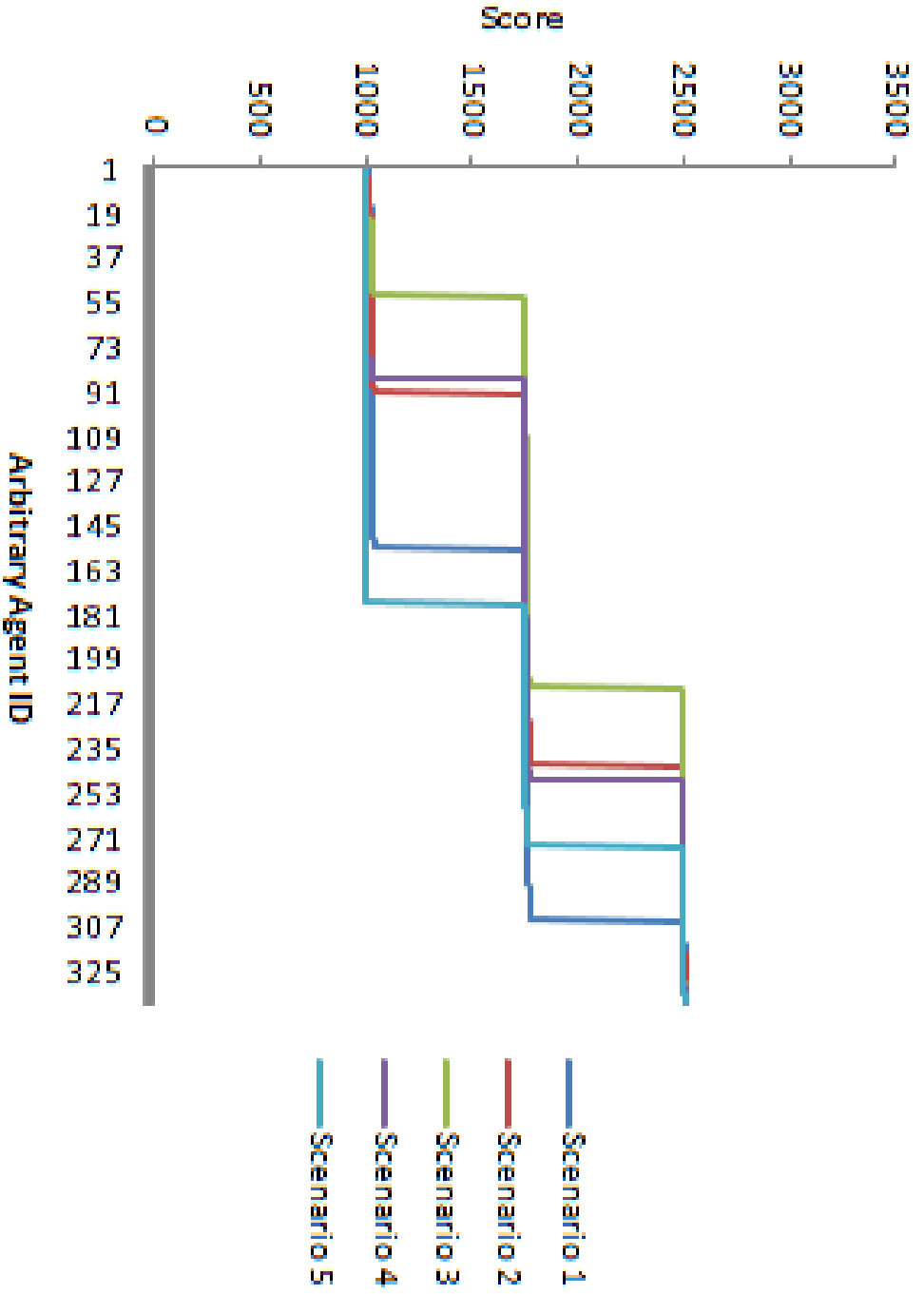

FIGURE A.5: Final scores of A2:G2:H2 agents from the first repeats of scenarios 1-5.

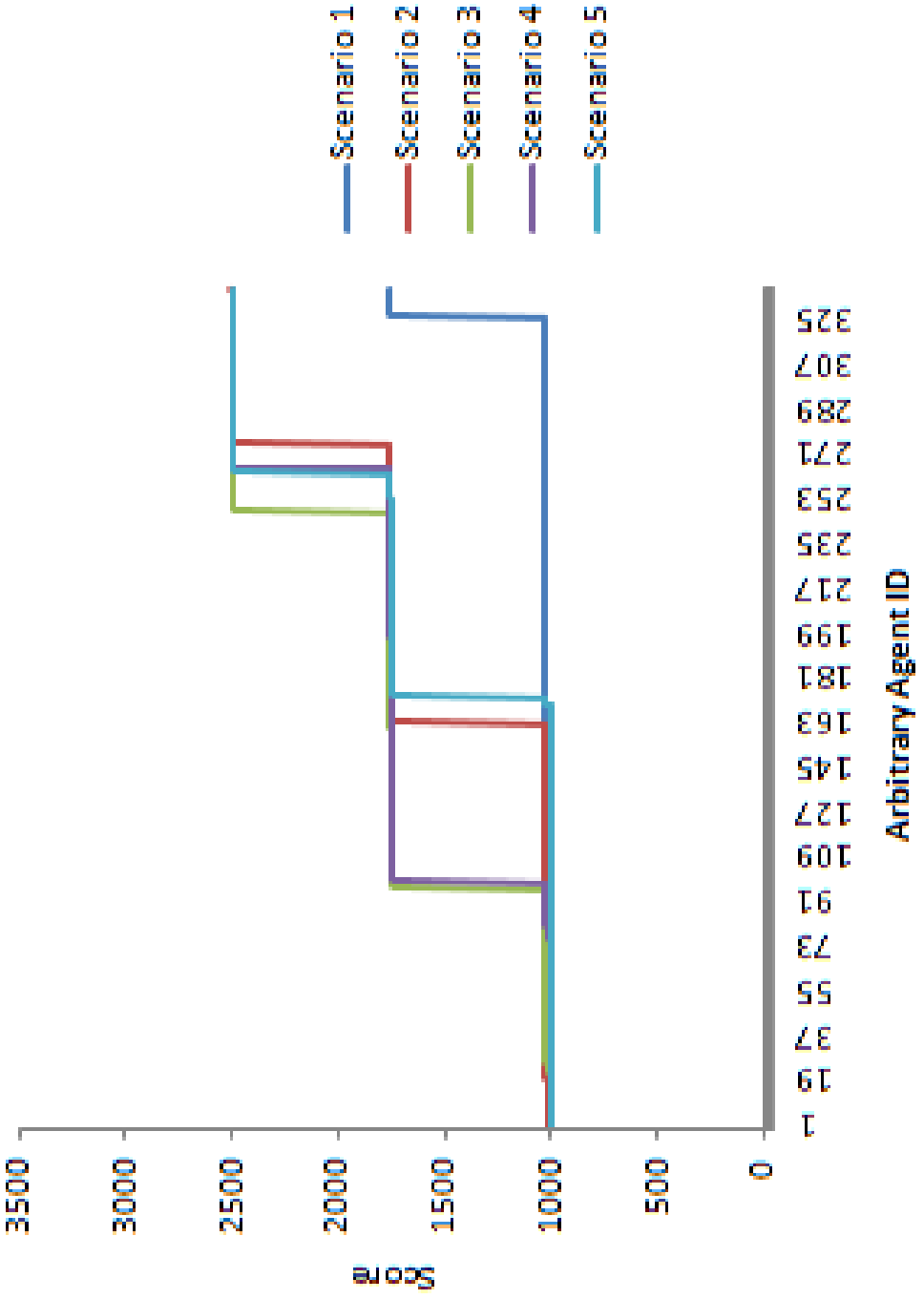

FIGURE A.6: Final scores of A2:G2:H3 agents from the first repeats of scenarios 1-5.

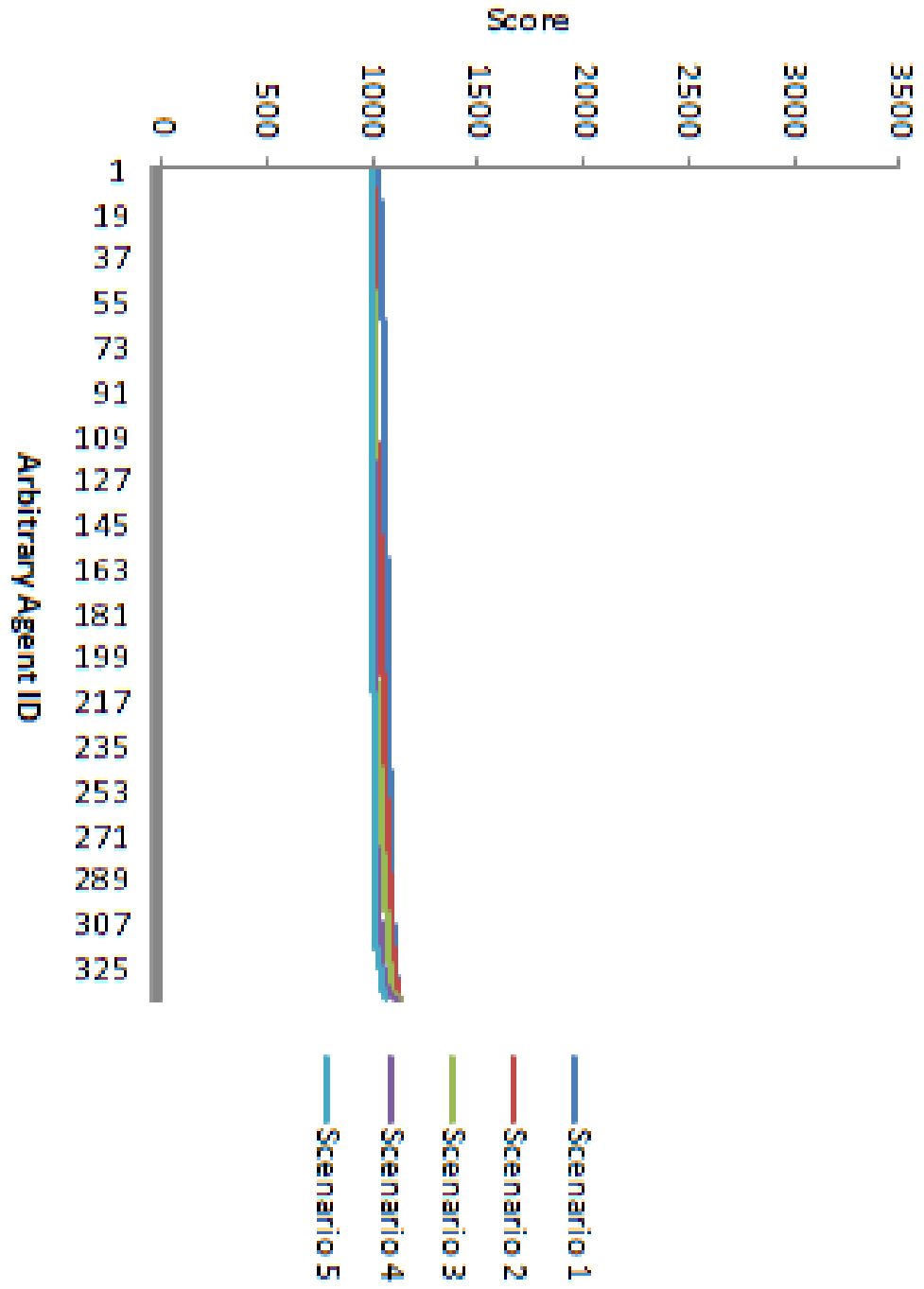

FIGURE A.7: Final scores of A2:G3:H1 agents from the first repeats of scenarios 1-5.

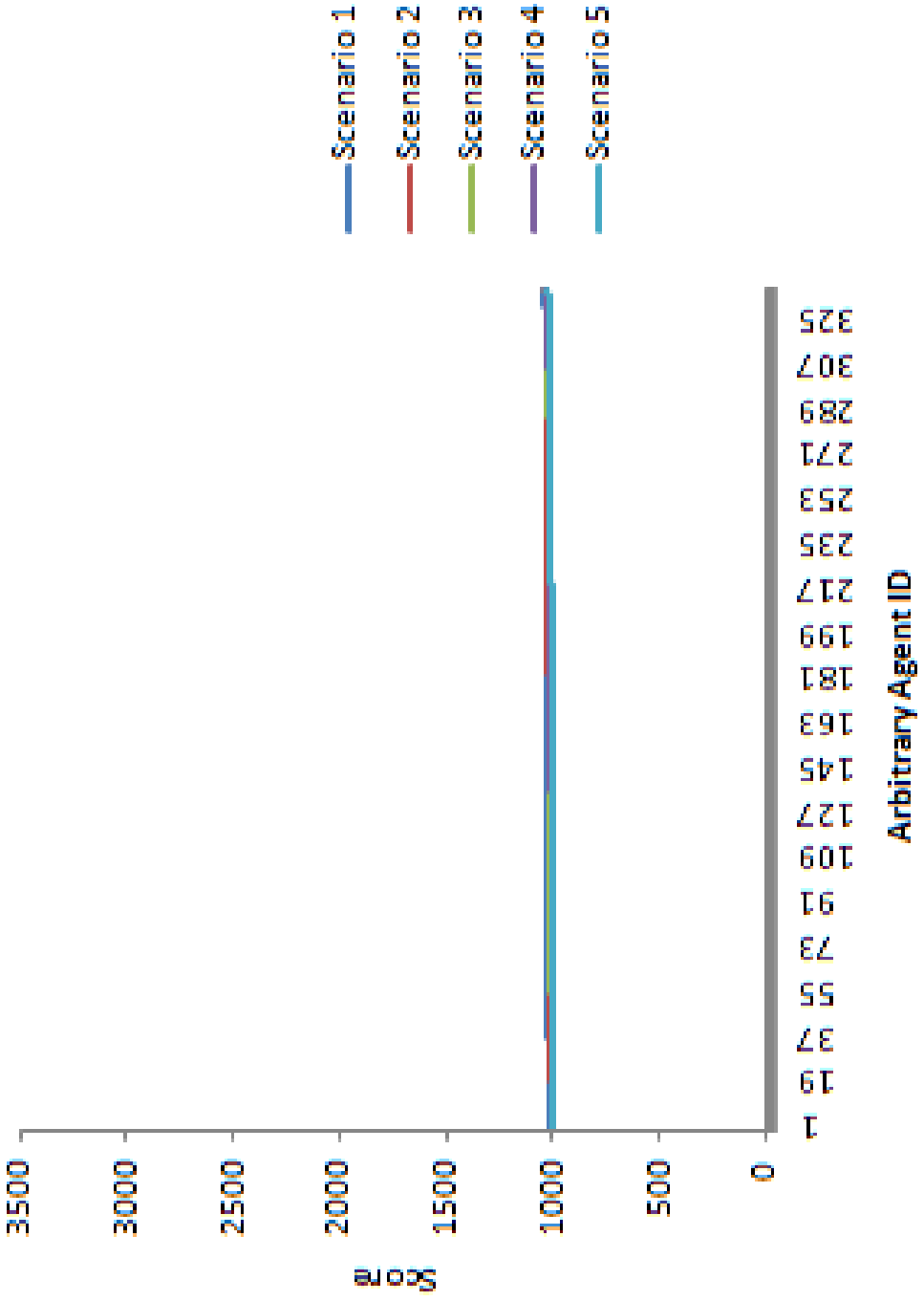

FIGURE A.8: Final scores of A2:G3:H2 agents from the first repeats of scenarios 1-5.

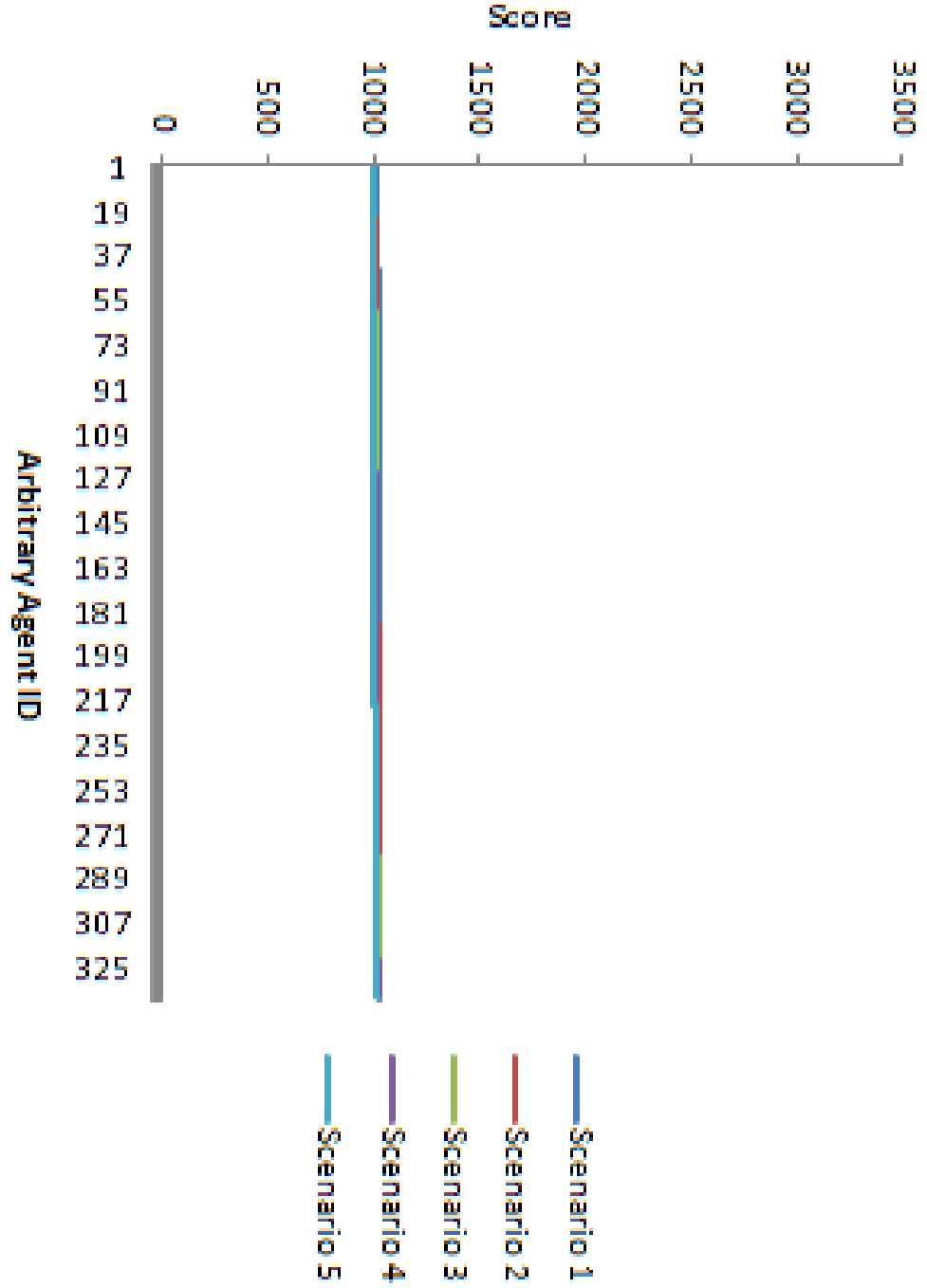

FIGURE A.9: Final scores of A2:G3:H3 agents from the first repeats of scenarios 1-5.

Appendix B

Analysis of how Emotional Character Types Determine Scores

The purpose of this appendix is to demonstrate that the conclusions of section 6.6 in chapter 6 and all results in chapters 7 and 8 are entirely deterministic and therefore, no significance testing of the results in chapters 6, 7 and 8 is required. Essentially, given the initial character type and initial behaviour of an emotional agent and its opponent(s), it is entirely possible to calculate the play history of the emotional agent and therefore, its final individual score without ever running the simulations detailed. However, given that a vast space of potential outcomes may emerge especially when there are many possible configurations, as in chapter 7, or, when simulations contain strategies capable of periodic defection (agents endowed with "hope", as implemented in chapter 8) it is much easier to build and run the simulations than it is to manually calculate the outcomes.

B.1 Emotional Agent Interaction Types

In the absence of strategies that periodically defect, interactions between agents with emotional characters always follow a course determined by the emotional characters of the agents involved and stabilise into either mutual cooperation, mutual defection or some kind of "*turn-taking*" play history for different periods (different combinations of emotional character types can alter periodicity). It is important to distinguish the different arrangements of turn-taking that may emerge between two agents and as such, there are two basic forms: symmetric and asymmetric turn-taking. Asymmetric turn-taking can be further decomposed into two types; sections B.1.1 and B.1.2 below discuss these types of turn-taking in more detail.

B.1.1 Symmetric Turn-Taking

Symmetric turn-taking occurs when, in round n, an emotional agent x cooperates with its opponent y who defects in round n. Then in round $n+1$, x switches to defection whilst y switches to cooperation. In round $n+2$, x switches back to cooperation whilst y switches back to defection. This switching back and forth between cooperation and defection every round occurs until the round limit is reached. Such a style of turn-taking is called $TT=$ since the agents involved finish the game with an equal individual score (provided an even number of rounds are played and agents establish this behaviour from round 1 onwards). Note also that the number of rounds between agents alternating their behaviour may be greater than one. For example, agent x may cooperate in rounds n and $n+1$ before switching to defection in rounds $n+2$ and $n+3$ and then cooperating once again in rounds $n+4$ and $n+5$ whilst y mirrors x's actions in each round.

This type of turn-taking is established between agents whose emotional characters are the same and contain anger and gratitude activation thresholds that are equal i.e. emotional agents A1:G1, A2:G2 and A3:G3. The maximum number of rounds before an agent switches its behaviour in symmetric turn-taking is three; this occurs when an initially cooperative A3:G3 agent plays against an A3:G3 agent that initially defects.

B.1.2 Asymmetric Turn-Taking

When asymmetric turn-taking is established between two agents, one agent in the pair will achieve a greater individual score by the conclusion of the game. For example, an emotional agent x in round n defects against its opponent y who cooperates, then, in round $n+1$, x defects again whilst y again cooperates. Finally, in round $n+2$, x switches to cooperation whilst y switches to defection resulting in x earning 10 points in total whilst y earns 5 points. In round $n+3$ both x and y switch back to their original behaviours: x defects for two rounds and cooperates for one whilst y cooperates for the two rounds in which x defects and defects in the one round where x cooperates. From x's perspective, this style of turn-taking is called $TT+$ since x earns 10 points every three rounds whereas from ys's perspective this style of turn-taking is called $TT-$ since y only achieves 5 points every three rounds. Note, however, that when compared to mutual defection, $TT-$ still imparts greater individual and total system scores for an agent, provided the agent administers at least one sucker's pay-off every four rounds.

Asymmetric turn-taking is achieved when the initial behaviour of two agents are different i.e. agent x's initial behaviour is set to cooperate whilst agent y's initial behaviour is set to defect. Also, the emotional characters of these agents must be set to the inverse of one another's i.e. if x's emotional character is set to A2:G1, y's must be set to A1:G2 (see table B.2).

B.2 Calculating Emotional Character Scores

In all the simulations run in this thesis, there are nine basic emotional characters and two potential initial behaviours (cooperate and defect). This produces 18 different emotional agent characters which can serve as opponents. For each emotional character it is possible to determine:

- The initial (pre-convergence) sequence of behaviour.
- The sequence converged on (mutual cooperation, mutual defection, symmetric turn-taking or asymmetric turn-taking).
- The expected score that the agent will attain expressed in terms of the number of rounds played, N.

This information acts as a foundation of validity for the results discussed in chapters 6, 7 and 8. Tables B.1 to B.9 contain the information described above for each of the 9 basic emotional characters. Note that each table is concerned predominantly with the initially cooperative version of the relevant emotional character, opponents all defect initially. To retrieve information for versions of emotional characters that initially defect, simply find its initially cooperative opponent's table, find the row that the emotional character that initially defects is contained within and read off the information from the relevant column.

Tables B.1 to B.9 do not show information pertaining to emotional characters with the same initial behaviour. Under such circumstances, the agents involved will simply continue to either mutually cooperate or defect depending upon their initial behaviour (assuming periodic defection is impossible). Consequently:

- All agents that cooperate initially and have initially cooperative opponents meet on CC in round 1 so each agent's expected score is $3N$.
- All agents that defect initially and have initially defective opponents meet on DD in round 1 so each agent's expected score is N.

For any emotional character that establishes $TT+$ or $TT-$ with its opponent, its opponent establishes the opposite e.g. in table B.2, when playing against A2:G1, A1:G2 converges upon $TT+$ however A2:G1 converges upon $TT-$.

Thus it is possible to calculate the expected score of any given emotional character in a population containing any specified mix of emotional characters by summing together the totals that each emotional character will earn when playing against each one of its opponents (the expected score of the agent in tables B.1 to B.9). This calculation is outlined in equation B.1 which produces the expected score for an agent of type n over r rounds where:

- w_i is equal to the percentage of type i agents divided by 100.

- s_{nir} is equal to the expected score that an agent of type n obtains when playing against an agent of type i over r rounds (can be read from the relevant table).

$$\sum_{1}^{18} w_i \times s_{nir} \tag{B.1}$$

Thus for example, suppose we have a homogeneous population of A1:G1 agents. The following statements are true:

- Agents in this population who initially cooperate will score:
  - $3N$ against initial cooperators.
  - $5N/2$ against initial defectors since $TT =$ occurs.
- Agents in this population who initially defect will score:
  - N against initial defectors.
  - $5N/2$ against initial cooperators since $TT =$ occurs.

It is then possible to calculate the expected scores for any agent given different percentages of initial cooperators and defectors in the homogeneous A1:G1 population. For example, if 100 rounds are played:

- If 50% of the population initially defect and 50% initially cooperate then:
  - Initial cooperators will have an expected score of $(0.5 \times 300) + (0.5 \times 250) = 275$.
  - Initial defectors will have an expected score of $(0.5 \times 100) + (0.5 \times 250) = 175$.
- If 20% of the population initially defect and 80% initially cooperate then:
  - Initial cooperators will have an expected score of $(0.8 \times 300) + (0.2 \times 250) = 290$.
  - Initial defectors will have an expected score of $(0.8 * 250) + (0.2 * 100) = 220$.
- If 80% of the population initially defect and 20% initially cooperate then:
  - Initial cooperators will have an expected score of $(0.2 * 300) + (0.8 * 250) = 260$.
  - Initial defectors will have an expected score of $(0.2 * 250) + (0.8 * 100) = 130$.

The requirement on the simulation is then to be sufficiently large for its results to tend towards these expected scores. "Large" in this sense refers to two facets of the simulation:

- The number of agents.

- The number of rounds.

If there are few agents involved there may be a disproportionate number of encounters between a particular pair of emotional characters which could skew the outcome of the simulation with respect to final individual scores. For example, if the number of A1:G1 and A2:G2 emotional characters in a population is set to 80% and 20% respectively then we can calculate the expected score of an A1:G3 agent using these percentages. However, if every agent in this simulation plays against 8 opponents then it may happen (by virtue of random agent placement in the initial set-up) that 90% of the A1:G3 agent's player set is composed of A1:G1 agents and 10% is composed of A2:G2 agents. This is obviously more likely to happen when the population size is relatively small than relatively large.

To allow for an individual agent's behaviour to be affected by its emotional character, the number of rounds in a simulation has to be sufficiently large. As a simple example, if we consider emotional characters A1:G1 and A3:G3, it can be seen that if the number of rounds in the simulation was set to 1 then neither emotional character would exert an effect upon the agent's behaviour. This means that the individual score of an agent is entirely determined by initial behaviour of the agents in question and the initial behaviour of their opponents. If the number of rounds in the simulation were set to 3 then an agent with the A1:G1 emotional character would have two rounds in which its emotional character has an opportunity to affect its behaviour. However, an A3:G3 agent's behaviour and therefore its individual score would be purely determined by its initial behaviour and the behaviour of its opponent. With respect to chapter 7, the number of rounds played should be sufficient to allow for every emotional character used in the simulations to exert an effect. In chapter 8, the number of rounds played should be sufficient to allow for hope's effect to be manifest i.e. for the actual results to approximate to the value expected in accordance with binomial probabilities. What is considered to be satisfactory will depend on how close to the expected value the simulation is required to be: while there will always be a difference between the expected and the actual values, this can be made as small as is desired by increasing the sample size.

B.3 Substantiating Thesis Results

The results of the expected score calculations (whose methodology is outlined in section B.2) can be used to substantiate claims made in the results of this thesis. For example, using the information in tables B.1, B.4 and B.7, the claim that increased tolerance enables mutual cooperation and thus promotes total system scores at the cost of individual scores (see chapter 6) can be validated. Indeed, table B.1 shows that an initially cooperative A1:G1 agent establishes mutual cooperation with 11/18 emotional characters, whereas A2:G1 (see table B.4) establishes mutual cooperation with 14/18 emotional characters and A3:G1 (see table B.7) with 17 emotional characters.

With respect to individual scores, mutual cooperation can only be beaten in the context of three rounds if an agent and its opponent converge upon $TT+$, and then only

if one of the agents can administer two sucker's pay-offs in return for one. An example of this can be seen when an initially cooperative A1:G3 emotional character plays against an A3:G1 emotional character that initially defects (see table B.3). However, as can be seen in tables B.2, B.3, B.4, B.6, B.7, B.8, $TT+$ encounters require very specific emotional characters in the player/opponent relationship. In a heterogeneous emotional character population where emotional characters are reasonably evenly spread, the large number of mutual defections these specific emotional characters fall into will considerably outweigh the few $TT+$ relationships.

When admiration is present within a population (see chapter 7), it is an emotional character's expected score at the end of 5 rounds that is important. The information in tables B.1 to B.9 can be used to determine which emotional characters can be expected to be admired by a given emotional character (if the initial set-up is known) and therefore the prevalence of each emotional character following each comparison. So, given a particular population mix, it is possible to determine the probability that one of the potentially admired emotional characters will be in the comparator set of a given agent. Consequently, the emotional character that the agent will emulate from its comparator set can then be determined. Again the validity of the simulation results depend on the size of the simulation being sufficiently large to allow the actual results to tend towards the expected scores.

For greed, the principles remains the same although, when two agents converge on mutual cooperation the following structure develops:

1. A period of mutual cooperation which always lasts for at least 3 rounds since hope's effect is only able to be manifest after 3 mutual cooperations with an opponent. The actual length of this step is determined by the probability of hope's effect being manifest in an agent. Given the probability of defection, this can be calculated using binomial probability (this is explained in the introduction to section 8.5 in chapter 8). If we assume that 100 opportunities for greed to take effect occur in a game then:

    - If the probability of defection for agents is 0.1 this will produce:
        - 81 mutual cooperations.
        - 9 cooperate-defects.
        - 9 defect-cooperates.
        - 1 mutual defection.
    - If the probability of defection for agents is 0.5 this will produce:
        - 25 mutual cooperations.
        - 25 cooperate-defects.
        - 25 defect-cooperates.
        - 25 mutual defection.
    - If the probability of defection for agents is 0.9 this will produce:

- 1 mutual cooperation.
- 9 cooperate-defects.
- 9 defect-cooperates.
- 81 mutual defections.

2. If hope's effect is manifest in an agent causing the agent to defect, we now have a situation which may be different from the initial situation. For example, if the two agents were both initial cooperators, these agents may now defect against each other or one may cooperate whilst the other defects.

3. Another convergence phase is then entered (which may or may not re-establish mutual cooperation) but which can be calculated from Tables B.1 to B.9 since the anger and gratitude thresholds of each agent and their first move of the new sequence are known.

   (a) If mutual cooperation is not re-established, the agents will continue in mutual defection, $TT =$ or $TT +/-$ for the remainder of the game.

   (b) If mutual cooperation is re-established go back to step 1 and continue from there.

As demonstrated above: given a probability of defection occurring after 3 mutual cooperations, binomial probability can be used to establish the probability of both, one or neither agent defecting. If the number of games is large enough, mutual cooperation will eventually break down into mutual defection. If the probability of defection is high, the probability of a mutual defection is correspondingly high and this will happen very quickly, but even with a low probability of defection it can be expected to happen eventually. This means that establishing mutual cooperation has its dangers, and that, in the long term, agents which establish turn-taking perform better than those which establish cooperation, since they do not risk mutual defection. This explains why agents whose threshold for anger and gratitude is equal are more robust in the face of greed's likelihood being high, as reported in chapter 8. The experiments in chapter 8 concerned homogeneous emotional character populations, but tables B.2, B.3, B.4, B.6, B.7, B.8 show which combinations establish a turn taking relationship and so will be resilient to greed.

B.4 Conclusion

Given the explanations in sections B.1, B.2 and B.3 it should be clear to see that it is possible to exactly calculate the expected individual score for every emotional agent in a population for any desired set-up, which can then be used to justify any claim made, without recourse to significance testing.

For example, if the results of a simulation suggest that A3:G1 outperforms character A2:G3 in a population containing all emotional characters with equal numbers of initial

defectors and cooperators, this can be validated absolutely by confirming this result from the expected values. A measure of confidence is not required, since we can know mathematically what the outcome will be. This accounts for the conclusions reported in section 6.6 of chapter 6.

With respect to chapter 7, it is possible to calculate the expected score of each emotional character at each comparison point, and the likelihood of an character being in an agent's comparator set. This information can then be used to produce the same results found in chapter 7 when running the simulations detailed in that chapter.

With respect to chapter 8, binomial probability is required to calculate the results obtained by running the simulations detailed in this chapter. Nonetheless, the same results would be produced so a result such as: "A2:G2 outperforms A1:G3 when hope is present in a population because in the long term, all emotional characters incapable of establishing symmetric or asymmetric turn-taking will converge upon mutual defection" can be substantiated using tables B.5 and B.3.

Obviously, running the simulations outlined in chapters 6, 7 and 8 gives a quicker and more visually pleasing way of demonstrating the effects (especially for admiration) and allowed for the generation of hypotheses. Moreover, provided that the simulation is large enough, the results of the simulation tend towards the expected results and this appendix demonstrates that this is so for the simulations used in this thesis. The validity of the claims can then be confirmed by the considerations advanced in this appendix.

TABLE B.1: Play history and scores of interest for an initially cooperative A1:G1 emotional character (player) when pitted against all other emotional character types that initially defect (opp.) for N rounds.

Emotional Character				**Score After 5 Rounds**		**Final Expected Score**	
Player	*Opp.*	**Converge On**	**Rounds to Converge**	*Player*	*Opp.*	*Player*	*Opp.*
A1:G1	A1:G1	TT=	0	10	15	5(N/2)	5(N/2)
	A1:G2	DD	1	4	9	N-5	N+5
	A1:G3	DD	1	4	9	N-5	N+5
	A2:G1	CC	2	14	14	3(N-2)+5	3(N-2)+5
	A2:G2	DD	1	4	9	N-5	N+5
	A2:G3	DD	1	4	9	N-5	N+5
	A3:G1	CC	2	14	14	3(N-2)+5	3(N-2)+5
	A3:G2	DD	1	4	9	N-5	N+5
	A3:G3	DD	1	4	9	N-5	N+5

TABLE B.2: Play history and scores of interest for an initially cooperative A1:G2 emotional character (player) when pitted against all other emotional character types that initially defect (opp.) for N rounds.

Emotional Character				**Score After 5 Rounds**		**Final Expected Score**	
Player	*Opp.*	**Converge On**	**Rounds to Converge**	*Player*	*Opp.*	*Player*	*Opp.*
A1:G2	A1:G1	DD	2	8	8	(N-2)+5	(N-2)+5
	A1:G2	DD	1	4	9	N-5	N+5
	A1:G3	DD	1	4	9	N-5	N+5
	A2:G1	TT+	0	15	10	10(N/3)	5(N/3)
	A2:G2	DD	1	4	9	N-5	N+5
	A2:G3	DD	1	4	9	N-5	N+5
	A3:G1	CC	3	16	11	3(N-3)+10	3(N-3)+5
	A3:G2	DD	1	4	9	N-5	N+5
	A3:G3	DD	1	4	9	N-5	N+5

TABLE B.3: Play history and scores of interest for an initially cooperative A1:G3 emotional character (player) when pitted against all other emotional character types that initially defect (opp.) for N rounds.

Emotional Character				**Score After 5 Rounds**		**Final Expected Score**	
Player	*Opp.*	**Converge On**	**Rounds to Converge**	*Player*	*Opp.*	*Player*	*Opp.*
A1:G3	A1:G1	DD	2	8	8	(N-2)+5	(N-2)+5
	A1:G2	DD	2	8	8	(N-2)+5	(N-2)+5
	A1:G3	DD	2	8	8	(N-2)+5	(N-2)+5
	A2:G1	DD	3	12	7	(N-3)+10	(N-3)+5
	A2:G2	DD	1	4	9	N-1	(N-1)+5
	A2:G3	DD	1	4	9	N-1	(N-1)+5
	A3:G1	TT+	0	15	10	15(N/4)	5(N/4)
	A3:G2	DD	1	4	9	N-1	(N-1)+5
	A3:G3	DD	1	4	9	N-1	(N-1)+5

TABLE B.4: Play history and scores of interest for an initially cooperative A2:G1 emotional character (player) when pitted against all other emotional character types that initially defect (opp.) for N rounds.

Emotional Character				**Score After 5 Rounds**		**Final Expected Score**	
Player	*Opp.*	**Converge On**	**Rounds to Converge**	*Player*	*Opp.*	*Player*	*Opp.*
A2:G1	A1:G1	CC	1	12	17	3(N-1)	3(N-1)+5
	A1:G2	TT-	0	5	10	5(N/3)	10(N/3)
	A1:G3	DD	2	3	13	N-2	(N-2)+10
	A2:G1	CC	1	12	17	3(N-1)	3(N-1)+5
	A2:G2	CC	3	11	16	3(N-3)+5	3(N-3)+10
	A2:G3	DD	2	3	13	N-2	(N-2)+10
	A3:G1	CC	1	12	17	3(N-1)	3(N-1)+5
	A3:G2	CC	3	11	16	3(N-3)+5	3(N-3)*10
	A3:G3	DD	2	3	13	N-2	(N-2)+10

TABLE B.5: Play history and scores of interest for an initially cooperative A2:G2 emotional character (player) when pitted against all other emotional character types that initially defect (opp.) for N rounds.

Emotional Character				**Score After 5 Rounds**		**Final Expected Score**	
Player	*Opp.*	**Converge On**	**Rounds to Converge**	*Player*	*Opp.*	*Player*	*Opp.*
A2:G2	A1:G1	CC	1	12	17	3(N-1)	3(N-1)+5
	A1:G2	DD	3	7	12	(N-3)+5	(N-3)+10
	A1:G3	DD	2	3	13	N-2	(N-2)+10
	A2:G1	CC	1	12	17	3(N-1)	3(N-1)+5
	A2:G2	TT=	0	10	15	10(N/4)	10(N/4)
	A2:G3	DD	2	3	13	N-2	(N-2)+10
	A3:G1	CC	1	12	17	3(N-1)	3(N-1)+5
	A3:G2	CC	4	13	13	3(N-4)+10	3(N-4)+10
	A3:G3	DD	2	3	13	N-2	(N-2)+10

TABLE B.6: Play history and scores of interest for an initially cooperative A2:G3 emotional character (player) when pitted against all other emotional character types that initially defect (opp.) for N rounds.

Emotional Character				**Score After 5 Rounds**		**Final Expected Score**	
Player	*Opp.*	**Converge On**	**Rounds to Converge**	*Player*	*Opp.*	*Player*	*Opp.*
A2:G3	A1:G1	CC	1	12	17	3(N-1)	3(N-1)+5
	A1:G2	DD	3	7	12	(N-3)+5	(N-3)+10
	A1:G3	DD	2	3	13	N-2	(N-2)+10
	A2:G1	CC	1	12	17	3(N-1)	3(N-1)+5
	A2:G2	DD	4	11	11	N-4	N-4
	A2:G3	DD	2	3	13	N-2	(N-2)+10
	A3:G1	CC	1	12	17	3(N-1)	3(N-1)+5
	A3:G2	TT+	0	15	10	15(N/5)	10(N/5)
	A3:G3	DD	2	3	13	N-2	(N-2)+10

TABLE B.7: Play history and scores of interest for an initially cooperative A3:G1 emotional character (player) when pitted against all other emotional character types that initially defect (opp.) for N rounds.

Emotional Character				**Score After 5 Rounds**		**Final Expected Score**	
Player	*Opp.*	**Converge On**	**Rounds to Converge**	*Player*	*Opp.*	*Player*	*Opp.*
A3:G1	A1:G1	CC	1	12	17	3(N-1)	3(N-1)+5
	A1:G2	CC	2	9	19	3(N-2)	3(N-2)+10
	A1:G3	TT-	0	5	20	5(N/4)	15(N/4)
	A2:G1	CC	1	12	17	3(N-1)	3(N-1)+5
	A2:G2	CC	2	9	19	3(N-2)	3(N-2)+10
	A2:G3	CC	4	8	18	3(N-4)+5	3(N-4)+15
	A3:G1	CC	1	12	17	3(N-1)	3(N-1)+5
	A3:G2	CC	2	9	19	3(N-2)	3(N-2)+10
	A3:G3	CC	4	8	18	3(N-4)+5	3(N-4)+15

TABLE B.8: Play history and scores of interest for an initially cooperative A3:G2 emotional character (player) when pitted against all other emotional character types that initially defect (opp.) for N rounds.

Emotional Character				**Score After 5 Rounds**		**Final Expected Score**	
Player	*Opp.*	**Converge On**	**Rounds to Converge**	*Player*	*Opp.*	*Player*	*Opp.*
A3:G2	A1:G1	CC	1	12	17	3(N-1)	3(N-1)+5
	A1:G2	CC	2	9	19	3(N-2)	3(N-2)+10
	A1:G3	DD	4	6	16	(N-4)+5	(N-4)+15
	A2:G1	CC	1	12	17	3(N-1)	3(N-1)+5
	A2:G2	CC	2	9	19	3(N-2)	3(N-2)+10
	A2:G3	TT-	0	10	15	10(N/5)	15(N/5)
	A3:G1	CC	1	12	17	3(N-1)	3(N-1)+5
	A3:G2	CC	2	9	19	3(N-2)	3(N-2)+10
	A3:G3	CC	5	10	15	3(N-5)+10	3(N-5)+15

TABLE B.9: Play history and scores of interest for an initially cooperative A3:G3 emotional character (player) when pitted against all other emotional character types that initially defect (opp.) for N rounds.

Emotional Character				**Score After 5 Rounds**		**Final Expected Score**	
Player	*Opp.*	**Converge On**	**Rounds to Converge**	*Player*	*Opp.*	*Player*	*Opp.*
A3:G3	A1:G1	CC	1	12	17	3(N-1)	3(N-1)+5
	A1:G2	CC	2	9	19	3(N-2)	3(N-2)+10
	A1:G3	DD	4	6	16	(N-4)+5	(N-4)+15
	A2:G1	CC	1	12	17	3(N-1)	3(N-1)+5
	A2:G2	CC	2	9	19	3(N-2)	3(N-2)+10
	A2:G3	DD	5	10	15	(N-5)+10	(N-5)+15
	A3:G1	CC	1	12	17	3(N-1)	3(N-1)+5
	A3:G2	CC	2	9	19	3(N-2)	3(N-2)+10
	A3:G3	TT=	0	10	15	15(N/6)	15(N/6)

Bibliography

[1] Robert P. Abelson and Vello Sermat, *Multidimensional scaling of facial expresions*, Journal of Experimental Psychology **63** (1962), no. 6, 546–554.

[2] Carole Adam, Andreas Herzig, and Dominique Longin, *A logical formalization of the OCC theory of emotions*, Synthese **168** (2009), no. 2, 201–248.

[3] David Aderman, *Elation, depression, and helping behaviour*, Journal of Personality and Social Psychology **24** (1972), no. 1, 91–101.

[4] T. K. Ahn, Elinor Ostrom, David Schmidt, Robert Shupp, and James Walker, *Cooperation in PD games: Fear, greed, and history of play*, Public Choice **106** (2001), no. 1-2, 137–155.

[5] Sara B. Algoe and Jonathan Haidt, *Witnessing excellence in action: the "other-praising" emotions of elevation, gratitude and admiration.*, The Journal of Positive Psychology **4** (2009), no. 2, 105–127.

[6] Justin Manuel Aronfreed, *Conduct and conscience*, Academic Press N. Y., 1968.

[7] Robert Axelrod, *The evolution of cooperation*, Basic Books, Inc., 1984.

[8] Albert Bandura, *Social cognitive theory: An agentic perspective*, Annual Review of Psychology **52** (2001), 1–26.

[9] Monica Y. Bartlett and David DeSteno, *Gratitude and prosocial behavior: Helping when it costs you.*, Psychological Science **17** (2006), no. 4, 319–325.

[10] Roy F. Baumeister, C. Nathan DeWall, Kathleen D. Vohs, and Jessica L. Alquist, *Then a miracle occurs: Focusing on behavior in social psychological theory and research*, ch. Does Emotion Cause Behavior (Apart from Making People Do Stupid, Destructive Things)?, pp. 119–136, Oxford University Press, 2009.

[11] Roy F. Baumeister, Kathleen D. Vohs, C. Nathan DeWall, and Liqing Zhang, *How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation*, Personality and Social Psychology Review **11** (2007), no. 2, 167–203.

[12] Ana L. C. Bazzan and Rafael H. Bordini, *A framework for the simulation of agents with emotions: Report on experiments with the Iterated Prisoners Dilemma*, Proceedings of the Fifth International Conference on Autonomous Agents, 2001.

[13] Antoine Bechara, *The role of emotion in decision-making: Evidence from neurological patients with orbitofrontal damage*, Brain and Cognition **55** (2004), 30–40.

[14] Antoine Bechara, Hanna Damasio, and Antonio R. Damasio, *Emotion, decision-making and the orbitofrontal cortex*, Cerebral Cortex **10** (2000), 295–307.

[15] Antoine Bechara, Daniel Tranel, and Hanna Damasio, *Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions*, Brain **123** (2000), 2189–2202.

[16] Trevor Bench-Capon, Katie Atkinson, and Peter McBurney, *Using argumentation to model agent decision making in economic experiments*, Autonomous Agents and Multi-Agent Systems **25** (2012), no. 1, 183–208.

[17] Trevor J. M. Bench-Capon, *Persuasion in practical argument using value-based argumentation frameworks*, Journal of Logic and Computation **13** (2003), no. 3, 429–448.

[18] Joyce Berg, John Dickhaut, and Kevin McCabe, *Trust, reciprocity and social history*, Games and Economic Behaviour **10** (1995), 122–142.

[19] Leonard Berkowitz and Philip Friedman, *Some social class differences in helping behavior*, Journal of Personality and Social Psychology **5** (1967), no. 2, 217–225.

[20] Ken Binmore, *Fun and games: A text on game theory*, D. C. Heath and Company: Lexington, MA, 1992.

[21] Ronald Bosman and Frans van Winden, *Emotional hazard in a power-to-take experiment*, The Economic Journal **112** (2002), 147–169.

[22] Simon Bowers and Rajeev Syal, *MP on Google tax avoidance scheme: 'I think that you do evil'*, `http://www.guardian.co.uk/technology/2013/may/16/google-told-by-mp-you-do-do-evil`, May 2013, Date Accessed: 17/05/2013.

[23] Samuel Bowles and Herbert Gintis, *Social capital and commiunity governance*, The Economic Journal **112** (2002), no. 483, F419–F436.

[24] Michael E. Bratman, *Intentions, plans, and practical reason*, Harvard University Press, 1987.

[25] Gertjan J. Burghouts, Dirk Heylen, Mannes Poel, Rieks Op Den Akker, and Anton Nijholt, *An action selection architecture for an emotional agent*, In Recent Advances in Artificial Intelligence, Proceedings of FLAIRS 16, Menlo Park: AAAI Press, 2003, pp. 293–297.

[26] Maxwell N. Burton-Chellew, Adin Ross-Gillespie, and Stuart A. West, *Cooperation in humans: Competition between groups and proximate emotions*, Evolution and Human Behavior **31** (2010), no. 2, 104–108.

[27] Lisa A. Cameron, *Raising the stakes in the Ultimatum game: Experimental evidence from Indonesia*, Economic Inquiry **37** (1999), no. 1, 47–59.

[28] Walter B. Canon, *The James-Lange theory of emotions: A critical examination and an alternative theory*, The American Journal of Psychology **39** (1927), no. 1-4, 106–124.

[29] Gary Charness and Matthew Rabin, *Understanding social preferences with simple tests*, The Quarterly Journal of Economics **117** (2002), no. 3, 817–869.

[30] Robert B. Cialdini, Betty Lee Darby, and Joyce E. Vincent, *Transgression and altruism: A case for hedonism*, Journal of Experimental Social Psychology **9** (1973), no. 6, 502–516.

[31] Kenneth Clark and Martin Sefton, *The sequential Prisoner's Dilemma: Evidence on reciprocation*, The Economic Journal **111** (2001), no. 468, 51–68.

[32] David J. Cooper and E. Glenn Dutcher, *The dynamics of responder behavior in Ultimatum games: A meta-study*, Experimental Economics **14** (2011), 519–546.

[33] Michael R. Cunningham, Jeff Steinberg, and Rita Grev, *Wanting to and having to help: Seperate motivations for positive mood and guilt-induced helping*, Journal of Personality and Social Psychology **38** (1980), no. 2, 181–192.

[34] Antonio Damasio, *Looking for Spinoza: Joy, sorrow, and the feeling brain*, Houghton Mifflin Harcourt, 2003.

[35] ———, *Descartes' error: Emotion, reason and the human brain*, Penguin, 2005.

[36] Mehdi Dastani and John-Jules Ch. Meyer, *Programming agents with emotions*, ECAI, 2006, pp. 215–219.

[37] Christopher T. Dawes, James H. Fowler, Tim Johnson, Richard McElreath, and Oleg Smirnov, *Egalitarian motives in humans*, Nature **446** (2007), 794–796.

[38] Robyn M. Dawes and Richard H. Thaler, *Anomalies; cooperation*, The Journal of Economic Perspectives **2** (1988), no. 3, 187–197.

[39] Ilona E. de Hooge, Seger M. Breugelmans, and Marcel Zeelenberg, *Not so ugly after all: When shame acts as a commitment device*, Journal of Personality and Social Psychology **95** (2008), no. 4, 933–943.

[40] Dominique J.-F. de Quervain, Urs Fischbacher, Valerie Treyer, Melanie Schellhammer, Ulrich Schnyder, Alfred Buck, and Ernst Fehr, *The neural basis of altruistic punishment*, Science **305** (2004), 1254–1258.

[41] Daniel C. Dennett, *The intentional stance*, The Massachussetts Institute of Technology, 1989.

[42] René Descartes, *The passions of the soul*, Hackett Publishing Company, 1989, Translator: Stephen Voss.

[43] David DeSteno, Monica Y. Bartlett, Jolie Baumann, Lisa A. Williams, and Leah Dickens, *Gratitude as moral sentiment: Emotion-guided cooperation in economic exchange*, Emotion **10** (2010), no. 2, 289–293.

[44] Paul Ekman, *An argument for basic emotions*, Cognition and Emotion **6** (1992), no. 3/4, 169–200.

[45] Paul Ekman, Wallace V. Friesen, Maureen O'Sullivan, Irene Diacoyanni-Tarlatzis, Rainer Krause, Tom Pitcairn, Klaus Scherer, Anthony Chan, Karl Heider, William Ayham LeCompte, Pio E. Ricci-Bitti, Masatoshi Tomita, and Athanase Tzavaras, *Universals and cultural differences in the judgments of facial expressions of emotion*, Journal of Personality and Social Psychology **53** (1987), no. 4, 712–717.

[46] Clark Elliott, *The Affective Reasoner: A process model of emotions in a multi-agent system*, Ph.D. thesis, Northwestern University, 1992.

[47] ———, *Using the Affective Reasoner to support social simulations*, IJCAI, 1993, pp. 194–201.

[48] ———, *Hunting for the holy grail with "emotionally intelligent" virtual actors*, SIGART Bulletin **9** (1998), no. 1, 20–28.

[49] Clark Elliott and Greg Siegle, *Variables influencing the intensity of simulated affective states*, Tech. Report Spring Symposium-93-05, AAAI Press, 1993.

[50] Jon Elster, *Emotions and economic theory*, Journal of Economic Literature **36** (1998), 47–74.

[51] ———, *Feelings and emotions: The amsterdam symposium (studies in emotion and social interaction)*, ch. Emotions and Rationality, pp. 30–48, Cambridge University Press, 2004.

[52] Robert N. Emde and David Oppenheim, *Self-conscious emotions: The psychology of shame, guilt, embarrassment, and pride*, ch. Shame, guilt, and the oedipal drama: Developmental considerations concerning morality and the referencing of critical others, pp. 413–438, New York: Guilford, 1995.

[53] Robert A. Emmons and Michael E. McCullough, *Counting blessings versus burdens: An experimental investigation of gratitude and subjective well-being in daily life*, Journal of Personality and Social Psychology **84** (2003), no. 2, 377–389.

[54] Christoph Engel, *Dictator games: A meta-study*, Experimental Economics **14** (2011), no. 4, 583–610.

[55] Trygg Engen, Nissim Levy, and Harold Schlosberg, *A new series of facial expressions*, American Psychologist **12** (1957), no. 5, 264–266.

[56] Ernst Fehr and Simon Gächter, *Cooperation and punishment in public goods experiments*, American Economic Review **90** (2000), no. 4, 980–994.

[57] ———, *Fairness and retaliation: The economics of reciprocity*, Journal of Economic Perspectives **14** (2000), no. 3, 159–181.

[58] ———, *Altruistic punishment in humans*, Nature **415** (2002), 137–140.

[59] Ernst Fehr and Klaus M. Schmidt, *A theory of fairness, competition, and cooperation*, The Quarterly Journal of Economics **114** (1999), no. 3, 817–868.

[60] Daniel M.T. Fessler and Kevin J. Haley, *Genetic and cultural evolution of cooperation*, ch. The Strategy of Affect: Emotions in Human Cooperation, pp. 7–36, MIT Press, 2003.

[61] Duncan Forest, Margaret S. Clark, Judson Mills, and Alice M. Isen, *Helping as a function of feeling state and nature of the helping behavior*, Motivation and Emotion **3** (1979), no. 2, 161–169.

[62] Robert Frank, *Passions within reason. the strategic role of the emotions*, New York: W. W. Norton & Company, 1988.

[63] Robert H. Frank, *If homo economicus could choose his own utility function, would he want one with a conscience?*, The American Economic Review **77** (1987), no. 4, 593–604.

[64] Robert H. Frank, Thomas Gilovich, and Dennis T. Regan, *Does studying economics inhibit cooperation?*, Journal of Economic Perspectives **7** (1993), no. 2, 159–171.

[65] ———, *The evolution of one-shot cooperation: An experiment*, Ethology and Sociobiology **14** (1993), 247–256.

[66] Robert W. Friedrichs, *Alter versus ego: An exploratory assessment of altruism*, American Sociological Review **25** (1960), no. 4, 496–508.

[67] Nico H. Frijda, *The emotions*, Cambridge University Press, 1986.

[68] Nico H. Frijda and Jaap Swagerman, *Can computers feel? Theory and design of an emotional system*, Cognition and Emotion **1** (1987), no. 3, 235–257.

[69] Dan Gardner, *Risk*, McClelland and Stewart Ltd, 2008.

[70] Jennifer M. George, *Emotions and leadership: The role of emotional intelligence*, Human Relations **53** (2000), 1027–1055.

[71] Thomas Gilovich, Dale Griffin, and Daniel Kahneman (eds.), *Heuristics and biases: The psychology of intuitive judgment*, Cambridge University Press, 2002.

[72] Corrado Gini, *Variabilitá e mutabilitá*, 1912.

[73] H. Gintis, *Beyond homo economicus: evidence from experimental economics*, Ecological Economics **35** (2000), no. 3, 311–322.

[74] Piotr J. Gmytrasiewicz and Christine L. Lisetti, *Emotions and personality in agent design and modeling*, Intelligent Agents VIII (John-Jules Ch. Meyer and Milind Tambe, eds.), Lecture Notes in Computer Science, vol. 2333, Springer Berlin Heidelberg, 2002, pp. 21–31.

[75] Carl Goldberg, *Understanding shame*, Northvale, NJ: Aronson, 1991.

[76] Julie H. Goldberg, Jennifer S. Lerner, and Philip E. Tetlock, *Rage and reason: The psychology of the intuitive prosecutor*, European Journal of Social Psychology **29** (1999), no. 5/6, 781–795.

[77] Alvin W. Gouldner, *The norm of reciprocity: A preliminary statement*, American Sociological Review **25** (1960), no. 2, 161–178.

[78] Dee L. R. Graham, Edna Rawlings, and Nelly Rimini, *Feminist perspectives on wife abuse*, ch. Survivors of terror: Battered women, hostages, and the Stockholm Syndrome, pp. 217–233, Sage Publications, 1988.

[79] Francesco Guala and Luigi Mittone, *Paradigmatic experiments: The Dictator game*, The Journal of Socio-Economics **39** (2010), no. 5, 578–584.

[80] Werner Güth, Rolf Schmittberger, and Bernd Schwarze, *An experimental analysis of ultimatum bargaining*, Journal of Economic Behaviour and Organization **3** (1982), 367–388.

[81] Werner Güth and Reinhard Tietz, *Ultimatum bargaining behavior: A survey and comparison of experimental results*, Journal of Economic Psychology **11** (1990), no. 3, 417–449.

[82] Jonathan Haidt, *Handbook of affective sciences*, ch. The Moral Emotions, pp. 852–870, Oxford: Oxford University Press, 2003.

[83] Garrett Hardin, *The Tragedy of the Commons*, Science **162** (1968), 1243–1248.

[84] Fritz Heider, *The psychology of interpersonal relations*, Wiley N. Y, 1958.

[85] Harry Helson, *Adaptation-level theory: An experimental and systematic approach to behavior*, Harper and Row, New York, 1964.

[86] Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath, *In search of homo economicus: Behavioral experiments in 15 small-scale societies*, The American Economic Review **91** (2001), no. 2, 73–78.

[87] Joseph Henrich and Francisco J. Gil-White, *The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission*, Evolution and Human Behavior **22** (2001), 165–196.

[88] Andreas Herzig and Dominique Longin, *C&L intention revisited*, Proceedings of the 9th International Conference on Principles of Knowledge Representation and Reasoning (KR2004) (Didier Dubois, Christopher A. Welty, and Mary-Anne Williams, eds.), AAAI Press, 2004, pp. 527–535.

[89] Kim Hill, *Altruistic cooperation during foraging by the Ache, and the evolved human predisposition to cooperate*, Human Nature **13** (2002), no. 1, 105–128.

[90] Jack Hirshleifer, *The latest on the best: Essays in evolution and optimality*, ch. On the Emotions as Guarantors of Threats and Promises, pp. 307–326, MIT Press, 1987.

[91] David Hume, *Treatise of human nature*, CreateSpace Independent Publishing Platform, 2013.

[92] Alice M. Isen, *Success, failure, attention, and reaction to others: The warm glow of success*, Journal of Personality and Social Psychology **15** (1970), no. 4, 294–301.

[93] Alice M. Isen and Paula F. Levin, *Effect of feeling good on helping: Cookies and kindness*, Journal of Personality and Social Psychology **21** (1972), no. 3, 384–388.

[94] WIlliam James, *What is an emotion?*, Mind **9** (1884), 188–205.

[95] Hong Jiang, Jose M. Vidal, and Michael N. Huhns, *EBDI: An architecture for emotional agents*, Proceedings of the 6th Internationl Joint Conference on Autonomous Agents and Multi-Agent Systems, AAMAS '07, ACM Press, 2007, pp. 38–40.

[96] Adam Gifford Jr., *Emotion and self-control*, Journal of Economic Behaviour and Organization **49** (2000), 113–130.

[97] Daniel Kahneman, *New challenges to the rationality assumption*, Journal of Institutional and Theoretical Economics (JITE) / Zeitschrift für die gesamte Staatswissenschaft **150** (1994), no. 1, 18–36.

[98] Daniel Kahneman, Jack L. Knetsch, and Richard H. Thaler, *Fairness and the assumptions of economics*, The Journal of Business **59** (1986), no. 4, S285–S300.

[99] ———, *Anomalies: The endowment effect, loss aversion, and status quo bias*, The Journal of Economic Perspectives **5** (1991), no. 1, 193–206.

[100] Daniel Kahneman, Paul Slovic, and Amos Tversky (eds.), *Judgement under uncertainty: Heuristics and biases*, Cambridge University Press, 1982.

[101] Hillard Kaplan, Kim Hill, Jane Lancaster, and A. Magdalena Hurtado, *A theory of human life history evolution: Diet, intelligence, and longevity*, Evolutionary Anthropolgy **9** (2000), 156–185.

[102] Dacher Keltner and James J. Gross, *Functional accounts of emotions*, Cognition and Emotion **13** (1999), no. 5, 467–480.

[103] Anthony Kenny, *Action, emotion and will*, Routledge, 1963.

[104] Timothy Ketelaar and Wing Tung Au, *The effects of feelings of guilt on the behaviour of uncooperative individuals in repeated social dilemma bargaining games: An affect-as-information interpretation of the role of emotion in social interaction*, Cognition and Emotion **17** (2003), no. 3, 429–453.

[105] Paul R. Kleinginna and Anne M. Kleinginna, *A categorized list of emotion definitions with suggestions for a consensual definition*, Motivation and Emotion **5** (1981), no. 4, 345–379.

[106] Yoram Kroll and Haim Levy, *Further tests of the seperation theorem and the capital asset pricing model*, The American Economic Review **82** (1992), no. 3, 664–670.

[107] Steven Kuhn and Serge Moresi, *Pure and utilitarian Prisoner's Dilemmas*, Economics and Philosophy **11** (1995), no. 2, 333–343.

[108] Richard Lazarus, *Cognition and motivation in emotion*, American Psychologist **46** (1991), no. 4, 352–367.

[109] Richard S. Lazarus, *Hope: An emotion and a vital coping resource against despair*, Social Research **66** (1999), no. 2, 653–678.

[110] Joseph Ledoux, *The emotional brain: The mysterious underpinnings of emotional life*, Simon and Schuster, 1998.

[111] Melvin J. Lerner and Rosemary R. Lichtman, *Effects of perceived norms on attitudes and altruistic behavior toward a dependent other*, Journal of Person **9** (1968), no. 3, 226–232.

[112] Martyn Lloyd-Kelly, Katie Atkinson, and Trevor Bench-Capon, *Developing co-operation through simulated emotional behaviour*, 13th International Workshop on Multi-Agent Based Simulation, MABS 2012, 2012.

[113] ———, *Emotion as an enabler of co-operation*, ICAART 2012 - Proceedings of the 4th International Conference on Agents and Artificial Intelligence, 2012, pp. 164–169.

[114] Penelope Lockwood and Ziva Kunda, *Superstars and me: Predicting the impact of role models on the self*, Journal of Personality and Social Psychology **73** (1997), no. 1, 91–103.

[115] ———, *Increasing the salience of one's best selves can undermine inspiration by outstanding role models*, Journal of Personality and Social Psychology **76** (1999), no. 2, 214–228.

[116] George Loewenstein, *Anticipation and the valuation of delayed consumption*, The Economic Journal **97** (1987), no. 387, 666–684.

[117] ———, *Out of control: Visceral influences on behavior*, Organizational Behavior and Human Decision Processes **65** (1996), no. 3, 272–292.

[118] George Loewenstein and Jennifer S. Lerner, *Handbook of affective science*, ch. The Role of Affect in Decision Making, pp. 619–642, New York: Oxford University Press, 2003.

[119] George Loewenstein and Nachum Sicherman, *Do workers prefer increasing wage profiles?*, Journal of Labor Economics **9** (1991), no. 1, 67–84.

[120] George Lowenstein, *Emotions in economic theory and economic behavior*, Preferences, Behavior and Welfare **90** (2000), no. 2, 426–432.

[121] Stacy Marsella, Johnathan Gratch, and Paolo Petta, *A blueprint for affective computing: A sourcebook and manual*, ch. Computational Models of Emotion, pp. 20–41, Oxford University Press, 2010.

[122] Lorna Marshall, *Sharing, talking and giving: Relief of social tension among !Kung bushmen*, Africa: Journal of the International African Institute **31** (1961), no. 3, 231–249.

[123] Benedetto De Martino, Dharshan Kumaran, Ben Seymour, and Raymond J. Dolan, *Frames, biases, and rational decision-making in the human brain*, Science **313** (2006), no. 5787, 684–687.

[124] Judith Masthoff and Albert Gatt, *In pursuit of satisfaction and the prevention of embarrassment: Affective state in group recommender systems*, User Modeling and User Adapted Interaction **16** (2006), no. 3-4, 281–319.

[125] Gerald Maxwell and Ruth E. Ames, *Economists free ride, does anyone else?*, Journal of Public Economics **15** (1981), no. 3, 295–310.

[126] Michael E. McCullough, Robert A. Emmons, and Jo-Ann Tsang, *The grateful disposition: A conceptual and empirical topography*, Journal of Personality and Social Psychology **82** (2002), no. 1, 112–127.

[127] Michael E. McCullough, Marcia B. Kimeldorf, and Adam D. Cohen, *An adaptation for altruism: The social causes, social effects, and social evolution of gratitude*, Current Directions in Psychological Science **17** (2008), no. 4, 281–285.

[128] Marvin Minsky, *The society of mind*, Simon & Schuster, 1988.

[129] Bert S. Moore, Bill Underwood, and D. L. Rosenhan, *Affect and altruism*, Developmental Psychology **8** (1973), no. 1, 99–104.

[130] J. Keith Murnighan and Michael Scott Saxon, *Ultimatum bargaining by children and adults*, Journal of Economic Psychology **19** (1998), no. 4, 415–445.

[131] Fahd Saud Nawwab, Trevor Bench-Capon, and Paul E. Dunne, *Emotions in rational decision making*, Argumentation in Multi-Agent Systems (ArgMAS 2009) (Peter McBurney, Iyad Rahwan, Simon Parsons, and Nicolas Maudet, eds.), Lecture Notes in Computer Science, vol. 6057, Springer, 2010, pp. 273–291.

[132] R. M. A. Nelissen, A. J. M. Dijker, and N. K. deVries, *How to turn a hawk into a dove and vice versa: Interactions between emotions and goals in a give-some dilemma game*, Journal of Experimental Social Psychology **43** (2007), no. 2, 280–286.

[133] Randolph M. Nesse, *The evolution of hope and despair*, Social Research **66** (1999), no. 2, 429–469.

[134] Martin Nowak and Karl Sigmund, *A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game*, Nature **364** (1993), 56–58.

[135] Martin A. Nowak and Robert M. May, *Evolutionary games and spatial chaos*, Nature **359** (1992), 826–829.

[136] Martin A. Nowak and Sébastien Roch, *Upstream reciprocity and the evolution of gratitude*, Proceedings of the Royal Society **274** (2007), 605–609.

[137] Martin A. Nowak and Karl Sigmund, *Tit-for-tat in heterogeneous populations*, Nature **355** (1992), 250–253.

[138] Edward O'Boyle, *The origins of homo economicus: A note*, Storia del Pensiero Economico **6** (2009), no. 1, 1–8.

[139] Fernando S. Oliveira, *Modeling emotions and reason in agent-based systems*, Computational Economics Vol. 35, Springer, 2009, pp. 155–164.

[140] Anthony J. Onwuegbuzie, *Role of hope in predicting anxiety about statistics*, Psychological Reports **82** (1998), 1315–1320.

[141] Anthony J. Onwuegbuzie and Carl R. Snyder, *Relations between hope and graduate students studying and test-taking strategies*, Psychological Reports **86** (2000), 803–806.

[142] Andrew Ortony, Gerald L. Clore, and Allan Collins, *The cognitive structure of emotions*, Cambridge University Press, 1988.

[143] Maffeo Pantaleoni, *Principii di economia pura*, Firenze, 1889.

[144] Philip Petit and Robert Sugden, *The backward induction paradox*, The Journal of Philosophy **86** (1989), no. 4, 169–182.

[145] Hans-Rüdiger Pfister and Gisela Böhm, *The multiplicity of emotions: A framework of emotional functions in decision making*, Judgment and Decision Making **3** (2008), no. 1, 5–17.

[146] Michel Tuan Pham, *Emotion and rationality: A critical review and interpretation of empirical evidence*, Review of General Psychology **11** (2007), no. 2, 155–178.

[147] Rosalind W. Picard, *Affective computing*, MIT Press, 1997.

[148] Madan M. Pillutla and J. Keith Murnighan, *Unfairness, anger, and spite: Emotional rejections of Ultimatum offers*, Organizational Behavior and Human Decision Processes **68** (1996), no. 3, 208–224.

[149] Mannes Poel, Rieks op den Akker, Anton Nijholt, and Aard-Jan van Kesteren, *Learning emotions in virtual environments*, In Proceedings of the 16th European meeting on cybernetics and systems research (Robert Trappl, ed.), vol. 2, 2002, pp. 751–756.

[150] Martha Pollack and Marc Ringuette, *Introducing the Tileworld: Experimentally evaluating agent architectures*, In Proceedings of the Eighth National Conference on Artificial Intelligence, AAAI Press, 1990, pp. 183–189.

[151] William Poundstone, *Prisoner's Dilemma*, Anchor, 1993.

[152] Vesna Prasnikar and Alvin E. Roth, *Considerations of fairness and strategy: Experimental data from sequential games*, The Quarterly Journal of Economics **107** (1992), no. 3, 865–888.

[153] Dean G. Pruitt, *Reciprocity and credit building in a laboratory dyad*, Journal of Personality and Social Psychology **8** (1968), no. 2, 143–147.

[154] Anatol Rapoport and Albert M. Chammah, *Prisoner's Dilemma*, University of Michigan Press, 1965.

[155] Scott Reilly, *Believable social and emotional agents*, Ph.D. thesis, Carnegie Mellon University (CMU), 1996.

[156] Raymond R. Reno, Robert B. Cialdini, and Carl A. Kallgren, *The transsituational influence of social norms*, Journal of Personality and Social Psychology **64** (1993), no. 1, 104–112.

[157] Milton E. Rosenbaum and Irving F. Tucker, *The competence of the model and the learning of imitation and nonimitation*, Journal of Experimental Psychology **63** (1962), no. 2, 183–190.

[158] Alvin E. Roth, *The handbook of experimental economics*, ch. Bargaining Experiments, pp. 253–348, Princeton: Princeton University Press, 1995.

[159] Alvin E. Roth, Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir, *Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study*, The American Economic Review **81** (1991), no. 5, 1068–1095.

[160] James A. Russell, *Core affect and the psychological construction of emotion*, Psychological Review **110** (2003), 145–172.

[161] Stuart Russell and Peter Norvig, *Artificial intelligence: A modern approach*, 3 ed., Prentice Hall, 2010.

[162] David Sally, *Conversation and cooperation in social dilemmas : A meta-analysis of experiments from 1958 to 1992*, Rationality and Society **7** (1995), no. 1, 58–92.

[163] William Samuelson and Richard Zeckhauser, *Status quo bias in decision making*, Journal of Risk and Uncertainty **1** (1988), no. 1, 7–59.

[164] Alan G. Sanfey, James K. Rilling, Jessica A. Aronson, Leigh E. Nystrom, and Jonathan D. Cohen, *The neural basis of economic decision-making in the Ultimatum game*, Science **300** (2003), 1755–1758.

[165] Jack Sawyer, *The altruism scale: A measure of co-operative, individualistic, and competitive interpersonal orientation*, American Journal of Sociology **71** (1966), no. 4, 407–416.

[166] Roger C. Schank and Robert P. Abelson, *Scripts, plans, goals and understanding*, Hillside, NJ: Erlbaum, 1977.

[167] Klaus Scherer, *Appraisal processes in emotion: Theory, methods, research*, ch. Appraisal considered as a process of multi-level sequential checking, pp. 92–120, New York and Oxford: Oxford University Press, 2001.

[168] Klaus R Scherer, *On the rationality of emotions: or, when are emotions rational?*, Social Science Information **50** (2011), no. 3-4, 330–350.

[169] Harold Schlosberg, *The description of facial expressions in terms of two dimensions*, Journal of Experimental Psychology **4** (1952), no. 4, 229–237.

[170] ———, *Three dimensions of emotion*, The Psychological Review **61** (1954), no. 2, 81–88.

[171] Amartya Sen, *The new palgrave: Utility and probability*, ch. Rational Behaviour, p. 198 216, New York-London: Norton, 1990.

[172] Mizuho Shinada, Toshio Yamagishi, and Yu Ohmura, *False friends are worse than bitter enemies: "altruistic" punishment of in-group members*, Evolution and Human Behavior **25** (2004), no. 6, 379–393.

[173] Karl Sigmund, *Punish or perish? Retaliation and collaboration among humans*, Trends in Ecology and Evolution **22** (2007), no. 11, 593–600.

[174] Herbert Alexander Simon, *Models of man: Social and rational - mathematical essays on rational human behavior in a social setting*, Wiley, 1957.

[175] Brent Simpson, *Sex, fear, and greed: A social dilemma analysis of gender and cooperation*, Social Forces **82** (2003), no. 1, 35–52.

[176] Robert Slonim and Alvin E. Roth, *Learning in high stakes Ultimatum games: An experiment in the Slovak Republic*, Econometrica **66** (1998), no. 3, 569–596.

[177] Paul Slovic, Baruch Fischhoff, and Sarah Lichtenstein, *Stevens' handbook of experimental psychology. vol 2: Learning and cognition*, ch. Decision Making, pp. 673–738, New York: Wiley, 1988.

[178] Deborah A. Small and George Lowenstein, *The devil you know: The effects of identifiability on punishment*, Journal of Behavioral Decision Making **18** (2005), no. 5, 311–318.

[179] Carl R. Snyder, *Hope, goal blocking thoughts, and test-related anxieties*, Psychological Reports **84** (1999), 206–208.

[180] Carl R. Snyder, Cheri Harris, John R. Anderson, Sharon A. Holleran, Lori M. Irving, Sandra T. Sigmon, Lauren Yoshinobu, June Gibb, Charyle Langelle, and Pat Harney, *The will and the ways: Development and validation of an individual-differences measure of hope*, Journal of Personality and Social Psychology **60** (1991), no. 4, 570–585.

[181] Carl R. Snyder, Susie C. Sympson, Florence C. Ybasco, Tyrone F. Borders, Michael A. Babyak, and Raymond L. Higgins, *Development and validation of the state hope scale*, Journal of Personality and Social Psychology **70** (1996), no. 2, 321–335.

[182] Charles R. Snyder, *Hope theory: Rainbows in the mind*, Psychological Inquiry **13** (2002), no. 4, 249–275.

[183] Sara J. Solnick and Maurice E. Schweitzer, *The influence of physical attractiveness and gender on ultimatum game decisions*, Organizational Behavior and Human Decision Processes **79** (1999), no. 3, 199–215.

[184] Robert C. Solomon, *A passion for justice: Emotions and the origins of the social contract*, Reading, MA: Addison-Wesley, 1990.

[185] Benedict De Spinoza, *The ethics*, Penguin Classics, 1996, Translator: Edwin Curley.

[186] Bas R. Steunebrink, Mehdi Dastani, and John-Jules Ch. Meyer, *A logic of emotions for intelligent agents*, 22nd Conference on Artificial Intelligence, AAAI Press, 2007, pp. 142–147.

[187] ———, *A formal model of emotions: Integrating qualitative and quantitative aspects*, European Conference on Artificial Intelligence, vol. 178, IOS Press, 2008, pp. 256–260.

[188] ———, *A formal model of emotion-based action tendency for intelligent agents*, Proceedings of the 14th Portuguese Conference on Artificial Intelligence (EPIA 09), 2009.

[189] ———, *The OCC model revisited*, Proceedings of the 4th Workshop on Emotion and Computing - Current Research and Future Impact, 2009.

[190] ———, *Emotions to control agent deliberation*, AAMAS, 2010, pp. 973–980.

[191] Paul G. Straub and J. Keith Murnighan, *An experimental investigation of Ultimatum games: Information, fairness, expectations, and lowest acceptable offers*, Journal of Economic Behaviour and Organization **27** (1995), no. 3, 345–364.

[192] Jacob Swagerman, *The artificial concern realization system ACRES: A computer model of emotion*, Ph.D. thesis, University of Amsterdam, 1987.

[193] Rajeev Syal and Patrick Wintour, *MPs attack Amazon, Google and Starbucks over tax avoidance*, `http://www.guardian.co.uk/business/2012/dec/03/amazon-google-starbucks-tax-avoidance`, December 2012, Date Accessed: 17/05/2013.

[194] György Szabó and Csaba Tőke, *Evolutionary Prisoners Dilemma game on a square lattice*, Physical Review E **58** (1998), no. 1, 69–73.

[195] Shelley E. Taylor and Marci Lobel, *Social comparison activity under threat: Downward evaluation and upward contacts*, Psychological Review **96** (1989), no. 4, 569–575.

[196] Abraham Tesser, Robert Gatewood, and Michael Driver, *Some determinants of gratitude*, Journal of Personality and Social Psychology **9** (1968), no. 3, 233–236.

[197] Richard Thaler, *Toward a positive theory of consumer choice*, Journal of Economic Behaviour and Organization **1** (1980), 39–60.

[198] Elizabeth Marshall Thomas, *The harmless people*, Random House, N. Y., 1958.

[199] Robert L. Trivers, *The evolution of reciprocal altruism*, The Quarterly Review of Biology **46** (1971), no. 1, 35–57.

[200] Edward Z. Tronick, *Emotions and emotional communication in infants*, American Psychologist **44** (1989), no. 2, 112–119.

[201] Jo-Ann Tsang, *Gratitude and prosocial behaviour: An experimental test of gratitude*, Cognition and Emotion **20** (2006), no. 1, 138–148.

[202] Amos Tversky and Daniel Kahneman, *The framing of decisions and the psychology of choice*, Science **211** (1981), no. 4481, 453–458.

[203] ———, *Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment*, Psychological Review **90** (1983), no. 4, 293–315.

[204] ———, *Loss aversion in riskless choice: A reference-dependent model*, The Quarterly Journal of Economics **106** (1991), no. 4, 1039–1061.

[205] Juan D. Velásquez, *Modeling emotions and other motivations in synthetic agents*, Proceedings of the 14th National Conference on Artificial Intelligence, AAAI Press, 1997, pp. 10–15.

[206] ———, *When robots weep: Emotional memories and decision-making*, Proceedings of the 15th National Conference on Artificial Intelligence, AAAI Press, 1998, pp. 70–75.

[207] Juan D. Velásquez and Pattie Maes, *Cathexis: A computational model of emotions*, Agents, 1997, pp. 518–519.

[208] Kathleen D. Vohs, Nicole L. Mead, and Miranda R. Goode, *The psychological consequences of money*, Science **314** (2006), 1154–1156.

[209] David Wechsler, *What consititues an emotion?*, Psychological Review **32** (1925), no. 3, 235–240.

[210] U. Wilensky, *Netlogo*, `http://ccl.northwestern.edu/netlogo`, 1999, Date Accessed: 23/6/2010.

[211] Ludwig Wittgenstein, *Philosophical investigations*, Blackwell Publishing, 1953.

[212] Michael Wooldridge, *An introduction to multi-agent systems*, 2nd ed., John Wiley & Sons, 2009.

[213] Michael Wooldridge and Nicholas R. Jennings, *Agent theories, architectures, and languages: A survey*, Intelligent Agents: ECAI-94 Workshop on Agent Theories, Architectures, and Languages (Michael Wooldridge and Nicholas R. Jennings, eds.), Lecture Notes in Computer Science, vol. 890, Springer, 1995, pp. 1–39.

[214] Erte Xiao and Daniel Houser, *Emotion expression in human punishment behavior*, Proceedings of the National Academy of Sciences of the United States of America **102** (2005), no. 20, 7398–7401.

[215] Marcel Zeelenberg, Rob M. A. Nelissen, Seger M. Breugelmans, and Rik Pieters, *On emotion specificity in decision making: Why feeling is for doing*, Judgment and Decision Making **3** (2008), no. 1, 18–27.

[216] Alexia Zoumpoulaki, Nikos Avradinis, and Spyros Vosinakis, *A multi-agent simulation framework for emergency evacuations incorporating personality and emotions*, Artificial Intelligence: Theories, Models and Applications (SETN 2010) (Joel A. Kubby, ed.), Lecture Notes in Computer Science, vol. 6040, 2010, pp. 423–428.