IN SEARCH OF SELF-CONTROL THROUGH COMPUTATIONAL MODELLING OF INTERNAL CONFLICT

Aristodemos Cleanthous

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Cyprus

Recommended for Acceptance

By the Department of Computer Science

October, 2010

2010

IN SEARCH OF SELF-CONTROL THROUGH COMPUTATIONAL MODELLING OF INTERNAL CONFLICT

Aristodemos Cleanthous

University of Cyprus, 2010

This thesis proposes a novel computational model of internal conflict which aims to provide further understanding on this highly complex and perplexing condition of the human brain. In particular, the purpose of this thesis is to identify specific factors which influence and enable internal conflict to be resolved by self-control behaviour.

Individuals are likely to experience an internal conflict when evaluating the same outcomes of choice along distinct dimensions or criteria. A value conflict of this sort can be resolved as if it was a result of strategic interaction between rational subagents of the brain. The particular setting for this interaction is a well-studied theoretical game, the Iterated Prisoner's Dilemma, where the mutual cooperation outcome of the game corresponds to the behaviour of self-control. The computational system developed for the purposes of this thesis realises this particular view of internal conflict by implementing two spiking neural networks as two agents competing in the Iterated Prisoner's Dilemma, where the agents pursue individual value maximisation through simultaneous but independent learning.

This high-level game theoretical approach to the problem of internal conflict incorporates at the same time biological realism through the employed neuronal model,

the process of learning, as well as by relating the agents and their actions in the game with particular brain regions and their functioning. In particular, the spiking neural networks comprise of leaky integrate-and-fire neurons, while the learning process is implemented by reinforcement of stochastic synaptic transmission as well as by reward-modulated spike-timing-dependent plasticity with eligibility trace. Moreover, the action of cooperation and defection by each agent maps to a greater relative activity of fronto-parietal and limbic system areas respectively.

As demonstrated through numerous simulations, the artificial neuronal system behaved efficiently in the game theoretical framework because the learning agents implemented the optimum result for the system through consistent mutual cooperation. Therefore self-control behaviour can indeed be learned (since it corresponds to mutual cooperation), and as showed by further results, it is enhanced by strong reward-correlated memory. Moreover, the ability of the agents to adopt optimal counter strategies as a response to their competitor's, enabled the identification of particular value structures that characterise internal conflicts of low and high intensity that promote or hinder the attainment of self-control behaviour.

In the process of obtaining the results which are relevant to the problem of self-control behaviour and internal conflict, this thesis work applied for the first time spiking neural agents combined with biological plausible reinforcement learning in a highly demanding multiagent task. In addition, further results with our system showed that high firing irregularity at high rates enhances learning.

APPROVAL PAGE

Doctor of Philosophy Dissertation

IN SEARCH OF SELF-CONTROL THROUGH COMPUTATIONAL MODELLING OF INTERNAL CONFLICT

Presented by

Aristodemos Cleanthous

Research Supervisor

_____

Chris Christodoulou

Committee Member

_____

Christos N. Schizas

Committee Member

_____

Edmund T. Rolls

Committee Member

_____

Guido Bugmann

Committee Member

_____

Marios N. Avraamides

University of Cyprus

October, 2010

Acknowledgements

I am heartily thankful to my supervisor, Dr Chris Christodoulou, for his guidance, support and confidence from the initial to the final stage of this research. I am also indebted to him for giving me the freedom to experiment with new ideas while at the same time keeping me on track. It would have been impossible to finish this thesis without his discreet but meaningful supervision. This work also owns a lot to Dr Gaye Banfield for setting the foundations on which this research is built. I am grateful to Dr Peter Sozou for his valuable comments as well as to Vassilis Vassiliades for our smooth collaboration. A special delayed thanks to Dr Denise Gorse for introducing me to the art of neural networks.

I gratefully acknowledge as well, the support of the University of Cyprus for a Small Size Internal Research Programme Grant and the Cyprus Research Promotion Foundation as well as the European Union Structural Funds for Grant PENEK/ENISX/0308/82.

I would also like to thank my beloved family for making this possible and my close friends who supported me all along this process, often with beers. Finally, I would like to thank Maria for everything.

Contents

List of Tables

List of Figures

List of Symbols

Symbols used by the learning algorithm of reward-modulated spike-timing-dependent plasticity with eligibility trace

w_{ij}: efficacy of the synapse from neuron j to i

γ: learning rate

δt: duration of a time step

r: global reward signal

z: eligibility trace

β: discount factor between 0 and 1

ζ: change of z resulting from the activity in the last time step

τ_z: is the time constant for the exponential decay of z

P^+ij: tracks the influences of presynaptic spikes

P^-_{ij}: tracks the influence of postsynaptic spikes

A_+: positive constant parameter

A_-: negative constant parameter

$f_i(t)$: signifies firing of neuron i

u_r : resting potential

θ: firing threshold

τ: time constant of membrane potential

Symbols used by the learning algorithm of reinforcement of stochastic synaptic transmission

p: probability of release of a neurotransmitter

q: release parameter

η: learning rate

h: global reinforcement signal

ē: eligibility trace

V_L: resting membrane potential

g_L: conductivity

C: capacitance

V_i: membrane potential of neuron i

E_{ij}: reversal potential of the synapse from neuron j to neuron i

G_{ij}: synaptic conductance

ΔG_{ij}: change in synaptic conductance

Wij: synaptic strength from neuron j to i

r_{ij}: neurotransmitter release variable

τ_s: synaptic conductance time constant

List of Abbreviations

C: Cooperate

D: Defect

CC: Cooperate-Cooperate

CD: Cooperate-Defect

DC: Defect-Cooperate

DD: Defect-Defect

STDP: Spike-Timing-Dependent Plasticity

PD: Prisoner's Dillema

IPD: Iterated Prisoner's Dilemma

LL: Larger-Later

SS: Smaller-Sooner

T: Temptation Payoff

R: Reward for mutual cooperation

S: Sucker's Payoff

P: Penalty for mutual defection

RL: Reinforcement Learning

MARL: Multi-Agent Reinforcement Learning

LIF: Leaky Integrate-and-Fire

XOR: Exclusive or

OLPOMDP: Online Partially Observable Markov Decision Process

Chapter 1

Introduction

Decisions are drawn and actions are performed on the conscious and the subconscious level, defining the perspective of ourselves to us and the ones around us. Understanding ourselves requires understanding the way we decide and act. Sometimes it is quite easy to make sense of our decisions and sometimes it is not. Many of our actions are executed out of pure biological need; we drink water, eat, sleep, make love etc., and although they sometimes need some serious planning before executed, no real choice is given to whether they should be executed or not. Additionally, quite often decisions are drawn based on pure preference ordering. For example I could easily tell why I choose beef over chicken or chicken over salad for dinner, as these choices depend on my personal taste preferences. If I were on diet however, the decision would not be that obvious. We regularly find ourselves in situations where we have to decide between alternatives which elicit conflicting preference orderings if evaluated according to different criteria. If I were on diet, I would still prefer the taste of beef, but at the same time I would prefer a fitter version of me. Such a situation involves an intra-personal conflict, making the decision more difficult and complex to make. Whether I give in to temptation by having the beef or exhibit self-control by eating a lower-calorie meal, I could equally justify my decision. But even if people are very good at justifying or making excuses for their decisions when

experiencing an internal conflict, it is quite obscure how they truly come to these decisions, which internal processes underlay their decisions and which factors influence them.

Internal conflicts are experienced by everyone on an everyday basis since individuals regularly find themselves in situations where they have to choose between an alternative with a higher overall value and a more tempting but ultimately inferior option. These contradicting alternatives elicit internal conflicts for the individual in so many different contexts. For example should I go out and have fun or stay home and finish some pending work? Or should I buy this cool gadget or save some money for the loan deposit? Should I order this delicious greasy burger or stick to a lower calorie meal in order to loose some weight? These are just a few examples of dilemmas faced by people among many. Of course different dilemmas apply to different people in quality and intensity. Not everybody needs to loose weight and even if he/she does, the intensity of the dilemma differs on whether he/she needs to loose 5kg or 20kg. In addition, the timing of the dilemma influences our choices as it is easier to resist temptation at one instant and very difficult at another. The fact is that we all experience internal conflicts at some point however small or big.

Maybe the most severe and challenging cases of internal conflict are presented during cases of addiction. The addicts experience full-blown internal conflicts, where decisions are often of life and death significance. However, the fact that in most cases addicts decide in favour of the short-term temptation not only challenges human rationality but at the same time raises questions on the power these individuals have on the outcome of such conflicts altogether. It is mind boggling to understand why smokers keep smoking even when they do not feel like smoking, knowing at the same time that

they consciously degrade their biological system. To smoke or not to smoke involves a dilemma, an internal conflict that shutters the brain, making the decision very difficult to make and the temptation to resist, and if not so, leaves the individual wondering about his/her decisions, actions and self.

No matter the intensity and the context of an internal conflict, individuals are faced with a dilemma between competing alternatives with different temporal values. Optimal decision making requires that the individual chooses the option which obtains the larger payoff in the long run; a behaviour known as self-control. In our examples, choosing to finish work, not to buy the cool gadget and eating the low fat meal would correspond to self-control behaviour as it goes along with the general definition of self-control which is to resist a smaller-sooner (SS) administered reward over a larger-later (LL) one (Rachlin, 1995). In other words, self-control would correspond to resist fun, unnecessary expenditure and food indulgence (examples of SS) over professionalism, better financial and physical fitness (examples of LL) respectively.

The current research thesis proposes a novel computational model of internal conflict. The model is designed after integrating relevant knowledge from such diverse areas such as psychology, game theory, cognitive and computational neuroscience. It aims at providing some further understanding on this highly complex and perplexing function of the human brain. Most importantly, the proposed model serves as a tool which is employed for the purposes of identifying the factors which influence and enable internal conflict to be resolved by self-control behaviour.

The presented work adopts the view that internal conflict can be resolved as if it was a result of strategic interaction between rational subagents of the brain (Kavka, 1991). The setting for this interaction is a well studied game, the iterated version of the

Prisoner's Dilemma (PD) (Rappoport and Chammah, 1965), whereas the Cooperate-Cooperate (*CC*) outcome of the game corresponds to the behaviour of self-control. The computational system developed for the purposes of this research thesis realizes this particular view of internal conflict by implementing two spiking neural networks as two agents competing in the iterated PD (IPD).

We implement this game theoretical view of internal conflict in a biologically and psychologically relevant computational model of spiking neural networks that learn through biologically plausible learning algorithms. Learning is implemented by reinforcement of stochastic synaptic transmission (Seung, 2003) as well as by reward-modulated spike-timing-dependent plasticity (STDP) with eligibility trace (Florian, 2007). The model does not intend to reproduce in detail the actual brain regions that are involved in an internal conflict. Too much information is missing with respect to the precise identity and function of the regions that enable and influence this highly complex state of mind. Given that, we believe that a high-level computational model which implements the brain as a functionally decomposed learning system, and involves internal conflicts as prescribed by game theory, would help us to acquire a better understanding of the big picture of internal conflict (and how it can be resolved), which is currently obscure. However, our computational model does not disregard important experimental findings (e.g. McClure et al., 2004) with respect to internal conflict and the brain areas involved. The identified brain regions and their functions were integrated in our game theoretical computational model by providing a plausible view of how the competing agents and these brain regions might relate. However no real subagents are presupposed. It is just helpful to explain and understand internal value conflicts as if they are represented by distinct rational internal agents.

Which factors influence and enable internal conflicts to be resolved by self-control behaviour? When is more likely for an individual to exhibit self-control? Why sometimes we excess on sweets while dieting does not begin on Mondays? These and similar questions will be pursued in search of self-control through computational modelling of internal conflict.

1.1 Thesis Outline

Chapter 2 reviews the literature on self-control and internal conflict as well as related aspects of decision making and intertemporal choice. It reviews relevant theoretical models of internal conflict and self-control with special emphasis on the model we implement. It presents related work and compares it with our approach to the computational modeling of internal conflict. The chapter ends by reviewing reinforcement learning on spiking neural networks.

Chapter 3 explains the design of the computational system with respect to its architecture and the IPD implementation. Moreover it presents the learning algorithms that are employed for the purposes of the research as well as their testing for correct implementation. It finishes by investigating the hypothesis that high firing irregularity at high rates enhances learning.

Chapter 4 presents simulations of internal conflict. It begins by describing how we overcame initial problems and continues by presenting results that demonstrate the ability of the system to exhibit self-control behaviour. Results are presented for both implemented learning algorithms. Moreover, it explores with additional simulations whether high firing irregularity at high rates enhances learning.

Chapter 5 investigates how the structure of internal value conflict influences the attainment of self-control behaviour. Experiments employ constant payoff structures that do not change during the duration of the game, as well as varying payoff structures, in order to simulate time related changes in the value systems of internal agents.

Finally, Chapter 6 overviews the computational model of internal conflict and the obtained results, lists the major contributions of this thesis work and suggests future directions.

Chapter 2

Literature Review

2.1 Theory of Self-Control, Internal Conflict and Intertemporal Choice

The self can be perceived as a goal directed hierarchical system, where goals are internally specified according to value systems (Scheier and Carver, 1988). However, the presence of more than one established value systems can divide the interest within a single individual and give rise to intra-personal conflict (Livnat and Pippenger, 2006). When experiencing an internal conflict, self-control behaviour can be employed and could be justified as one's desire to maximise long term reward (Barkley, 1997; Kanfer and Karoly, 1972; Mischel, 1996). Self-control is the exertion of control over the self by the self in order to prevent or inhibit its dominant response (Muraven et al., 1999). It is a behaviour during which a person is required to control his/her thoughts, emotions or actions that would otherwise automatically have or do. To exercise self-control requires to override or inhibit pre-existing automatic processes, which are quite robust and efficient and hence resistant to change (Muraven et al., 1999). The most essential feature of self-control behaviour is that it postpones immediate gratification in order to attain delayed but more valuable outcomes (Mischel et al., 1989). Rachlin (1995) also provided a broad definition of self-control based on that distinctive feature by saying that self-control can be defined as choosing a large delayed reward over a small immediate one.

Although this definition is quite general, it provides simplicity and consistency in defining a highly complex and perplexing behaviour as the majority of self-control problems can be translated into problems of delayed gratification. For the purposes of this thesis self-control is defined as such.

According to Ariely (2002), and Rachlin (2000), we recognise that we have self-control problems and try to solve them by precommitment behaviour. Precommitment behaviour can be seen as a desire by people to protect themselves against a future lack of willpower or as a strategy employed in order to avoid the experience of a subsequent internal conflict. Precommitment is more formally defined as making a choice now with the specific aim of denying (or at least restricting) oneself future choices (Rachlin, 1995). A typical example of precommitment is putting an alarm clock away from your bed, to force you to get up to turn it off. The effect of precommitment behaviour is the same as if self-control behaviour was exercised without the individual necessarily experiencing an internal conflict. Precommitment requires that people know which of the alternatives is best for them in the long run so that they precommit to the one with the highest payoff. According to experiments by Richmond et al. (2003), the brain's ability to recognise or predict future rewards is built in and past experience enhances this ability.

Internal conflict is manifested in many experiments of animal behaviour. For example, when a rat is offered both a reward (food) and penalty (electric shock) at a location, it oscillates at a certain distance from it, given certain parameters of reward and penalty (Miller, 1944). This oscillation is observed due to conflicting tendencies that co-exist at a dynamic equilibrium (Brown, 1948; Miller, 1944). The attribution of internal conflict to the rat implies the existence of two agents: one whose goal is to satisfy hunger and another whose goal is to avoid danger (Livnat and Pippenger, 2006). Animals also

express internal conflict in the form of ambivalence, exhibiting both aggression and courtship when approached by a female in their territory or by simultaneously peaking and incubating a painted egg that has been placed in their nest (Tinbergen, 1952). These results reveal that conflicting behaviours can co-exist and may be independent.

In humans, internal conflict is experienced on an everyday basis, in different contexts and intensities. Although it is very difficult to describe, we all know the annoying, unresting feeling when trying to decide between competing alternatives. In order to understand how people resolve such conflicts, scientists study decision making by employing either formal models like expected utility theory (von Neumann and Morgenstern, 1947) and prospect theory (Kahneman and Tversky, 1979) or reason-based models. Formal models make use of numerical values which are attached to the different alternatives and decision is characterised by value maximisation. In reason-based models, decision is justified according to the balance of reasons and arguments for and against the different alternatives. Although the two types of models do not contradict each other, different disciplines study decision making using one or the other type. For example, economic theory and management theory employ formal models in experimental studies of preference and in standard economic analyses whereas reason-based models explain decisions informally in the absence of experimental data as in business and law case-studies or in the interpretation of historic and political decisions.

However neither of the two models is sufficient to explain decisions in certain cases of internal conflict. For example it is often hard to assign values to alternatives, especially when involved in complex, real-life decision making. Moreover, even if values could be attached to the different alternatives, it would be impossible to reach a decision based on pure preference ordering in cases where more than one value system is in effect.

Food for instance can be valued either for the taste or nutritional value and these valuations could be quite contradicting. When both value systems are taken into account (e.g., when on diet) a choice of a certain food over another cannot be explained based on value ordering since an explanation should satisfy both value systems which are concurrently in effect. On the other hand, reason-based models have similar limitations in such cases as the analysis would result in contradicting reasons for competing options. In addition, people do not always know the actual reasons that guide their decisions and can come up with false explanations when asked to account for their decisions (Nisbett and Wilson, 1977), as every decision can be easily rationalised after it was taken.

Internal conflict has been mostly studied with respect to intertemporal choice tasks, where numerous experiments were conducted in order to understand decision making in cases where immediate and delayed rewards were competing for selection. Intertemporal choice is concerned with tradeoffs among outcomes occurring at different points in time (Frederick et al., 2002). Figure 2.1 illustrates a choice between a smaller-sooner (SS) reward, available at time t_2 and a larger-later (LL) reward, available at t_3. The ability to wait for the superior option is also known as delay of gratification (Mischel et al., 1989).

Future gains and losses are discounted when facing a choice between a smaller immediate gain or loss and a larger future one (Ainslie, 1975; Ainslie, 2001). The thin lines subtended from points SS and LL are temporal discount functions indicating the effect of increasing the delay of the reward.

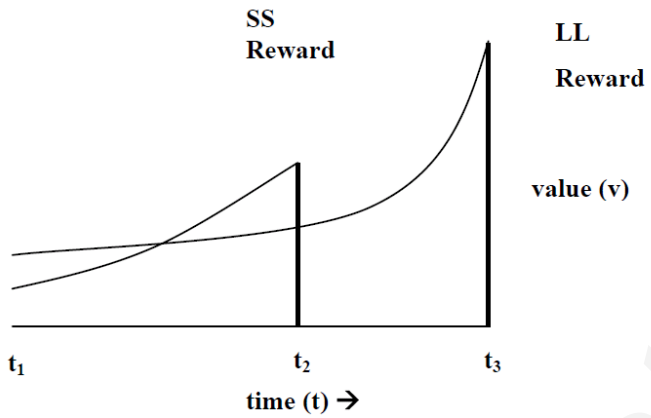

Figure 2.1: Delay of gratification

The figure illustrates a choice between a smaller-sooner (SS) reward, available at time t_2 and a larger-later (LL) reward, available at t_3. Notice that when SS is available at time t_2, its value is greater than the discounted value of LL. A person exhibits self control when s/he does not gather the available reward and waits for the bigger one (based upon Rachlin, 1995).

Several studies (Loewenstein and Prelec, 1992; Ainslie, 1975; Herrnstein, 1981; Mazur, 1987) have proposed hyperbolic functions in order to describe the phenomenon of temporal discounting, departing from the initial exponential function originally proposed by Samuelson (1937) in his 'discounting utility' model. Experiments have shown that the exponential model cannot account for their findings and most importantly for the decreasing discount rates, as observed in numerous studies (Thaler 1981; Benzion et al., 1989; Pender, 1996; Frederick et al., 2003). Results showed that discount rates decline as one looks further into the future, which also means that the discount function should flatten out more than the exponential. Or conversely as shown in Figure 2.1, the closer you get to a reward, the faster its current value increases. A consequence of the decreasing discount rates is the reversal of preferences which was also experimentally observed (Herrnstein 1990; Kirby and Herrnstein, 1995) and is depicted in Figure 2.1 by

the crossing of the hyperbolic discount functions at some point between t_1 and t_2. At time t_1, the value of the larger-later (LL) reward exceeds the smaller-sooner reward (SS). However at t_2 when SS would be immediately available, the value of SS exceeds LL value.

Self control behaviour is exhibited when one chooses LL even when SS is freely available and has a higher immediate value than LL. The possibility that LL might not be available in the future due to uncertainty does not change the analysis. The value of LL represents the expected value of the reward, taking into account the risk it entails. The case where the risk is so great such that at t_1 the value of LL becomes lower than the value of SS, does not concern us since the choice of SS is obvious.

In contrast to the exponential model, the hyperbolic model accounts both for the decreasing discount rates and the reversal of preferences. However none of the two models incorporates the plethora of experimental findings, which reveal some other important aspects of intertemporal choice, such as the absolute magnitude effect (Thaler 1981, Ainslie and Haendel 1983; Loewenstein, 1987; Benzion et al., 1989; Kirby and Marakovic, 1995; Kirby, 1997) and the gain loss asymmetry (Loewenstein 1987; Loewenstein and Prelec, 1992; Benzion et al.,1989).

Delay of gratification is the ability to wait for a delayed reward in the presence of an immediate inferior option. Mischel and his colleagues performed a series of experiments over the years (for review and findings reported below see Mischel et al., 1989) using the same subjects in order to investigate the psychological processes that enable delay of gratification and how the presence or absence of this ability in the early years of an individual's life correlates with future behaviour. When 4-year olds were offered to either wait for a preferred treat (candy or toy) or accept an inferior one

immediately or by terminating waiting time, findings surprisingly showed that attention to rewards either by exposure to the actual rewards during delay or by thinking about the rewards, consistently and substantially decreased the time they could wait. On the other hand, children managed to wait more through distractive 'fun' thoughts (as were instructed beforehand), by covering their eyes, by looking away from rewards and by other self-imposed distraction mechanisms. Despite that, attention to reward promoted delay of gratification when subjects were exposed to symbolic representations of the preferred reward (e.g., real size picture of the reward). In addition, focusing on abstract qualities and associations of the reward induced an increase in the waiting time as opposed to the case where the focus was on arousing qualities (e.g., taste of eating or playing with). Finally and surprisingly, one of the longest mean waiting times was observed when children were asked to think about the arousing qualities of comparable control objects (for example children waiting for marshmallows who had been cued to think about the salty, crunchy taste of pretzels) instead of their abstract qualities. Findings which support that children can be taught to suppress impatience by manipulation of thought were also presented in a later study (Metcalfe and Mischel, 1999). A recent study on students' delay of gratification (Bembenutty, 2009) also showed that the behaviour is enhanced with the use of self-regulated learning strategies like reminding themselves of their overall values and goals. In addition, the behaviour was accounted on "the relative value and expectation of success of engaging in delayed versus immediate activities typically faced by students". While commenting on the findings, Pychyl (2009) wrote: "To the extent that the students feel that they will succeed at a task that is valuable to them, they don't perceive the task as aversive (an emotional response, not an issue of utility per se) and approach the task rather than avoid it. If students find the task aversive

(typically because they feel a lack of competence or self-efficacy), their focus will be on short-term emotional repair, and they "give in to feel good" by engaging in the alternative task at the expense of their long-term goals". In the current thesis we believe that an emotional response is always evoked in an internal conflict task such as in delay of gratification, that can shape but also dynamically change the values (in terms of expected utility, as in Chapter 5) attached to the alternatives. The outcome of the behaviour cannot be solely attributed to either of the two processes, but instead should be accounted on the interplay of the two.

Recent research on the neural mechanisms that underlie intertemporal choice revealed that such behaviors result from competition between several neural networks of the human brain that interact with each other during a decision (McClure et al., 2004; McClure et al., 2007; Wittmann et al., 2007). A neuroimaging study by McClure et al. (2004) has shown that immediate rewards activate paralimbic areas, including the ventral striatum, medial orbitofrontal cortex, and medial prefrontal cortex whereas the lateral prefrontal cortex and posterior parietal cortex are engaged uniformly by intertemporal choices irrespective of delay. Furthermore, the relative engagement of the two systems is directly associated with subjects' choices, with greater relative fronto-parietal activity when subjects choose longer term options. Tanaka et al. (2004) also showed that choices of collecting immediate rewards activated lateral orbitofrontal cortex and striatum, whereas the dorsolateral prefrontal cortex and inferior parietal cortex were activated when subjects chose to obtain large future rewards. In general, these studies demonstrated that two separate neural systems are involved in such decisions; parts of the limbic system associated with the midbrain dopamine system, including paralimbic cortex, are preferentially activated by decisions involving immediately available rewards and fronto-

parietal areas are involved when choosing large future options. However, in contrast to the previous results, the subjective value of monetary rewards was shown to be represented by a single system (Kable and Glimcher, 2007), irrespective of delay upon delivery.

All the aforementioned studies investigated intertemporal choice with respect to immediate and future rewards and gains. When it comes to immediate and future losses, Xu et al. (2009) suggested that a common fronto-parietal network is used to discount both future gains and losses and its neural activity is stronger during loss discounting, indicating an asymmetric discounting with respect to gains and losses and a possible explanation of why future losses are discounted less steeply than future gains. In addition, the insula, thalamus and dorsal striatum were more activated during intertemporal choices involving losses, suggesting that the enhanced sensitivity to losses may be driven by negative emotions.

All the above neuroimaging studies (except maybe for the case of Kable and Glimcher, 2007) are consistent with the view that decision making incorporates several competing neural networks (De Martino et al., 2006; McClure et al., 2004b; Sanfey et al., 2003; Sanfey et al., 2006). Moreover, according to O'Reilly and Munakata (2000), the higher cognitive functions are not based on the action of individual neurons in a limited area, but on the outcome of the integrated action of the brain as a whole. Given the complexity of decision making with respect to internal conflict and the scope of this thesis, we chose to model the brain from a top-down perspective as a functionally decomposed system where attention is primarily focused on the competition between constructing modules. However, while the design of the system follows a top-down

approach, its function is defined and controlled by variables affected by learning in the neuronal level in a bottom-up approach.

2.2 Models of Self-Control

What follows is an overview of some dual-process models relevant to self-control as well as specific models designed to account for self-control behaviour. What is common in almost every model of self-control is that it involves a conflict. Adam Smith in his book "Theory of moral sentiments" (1759) describes a conflict between reason and passion, Thaler and Shefrin (1981) propose a two-self model of myopic doer *versus* far-sighted planner, Smolensky (1988) suggests a top-level conscious processor for effortful reasoning and an intuitive processor for intuitive problem solving whereas Christianity promotes a constant battle between good and evil. Whether self-control is a "fruit of the Holy Spirit" (Paul to Galatians (5:22-23)), is currently unknown, but the evidence of the duality that is associated with self-control is scientifically prominent.

Considering these models we could say that self-control is a certain way of resolving the conflict in the presence of two processes or 'entities' and more specifically in a way that promotes future well being for the individual as opposed to immediate gratification. However both of these processes could be employed for the purposes of either outcome. For example consider individuals like addicts who often engage reason in order to construct elaborate plans for acquiring their dose. In such a case, the process that corresponds to reason, the far-sighted planner or the top-level conscious processor is working towards immediate gratification.

Carver and Scheier (1998; 1982) propose an abstract model of self-control based on self-regulation of behaviour through feedback. Self-control corresponds to the operate

phase of self-regulation, where the operate phase refers to any sort of action that seeks to reduce discrepancies between a perceived aspect of self and a standard. To exercise self-control is thus to change the self in order to maintain conformity to a standard. The model is diagrammatically shown in Figure 2.2. Although the model is quite general and can be applied in various self-control contexts, it lacks neurobiological realism as it does not justify the existence, or provides a correspondence of its components in the real brain.

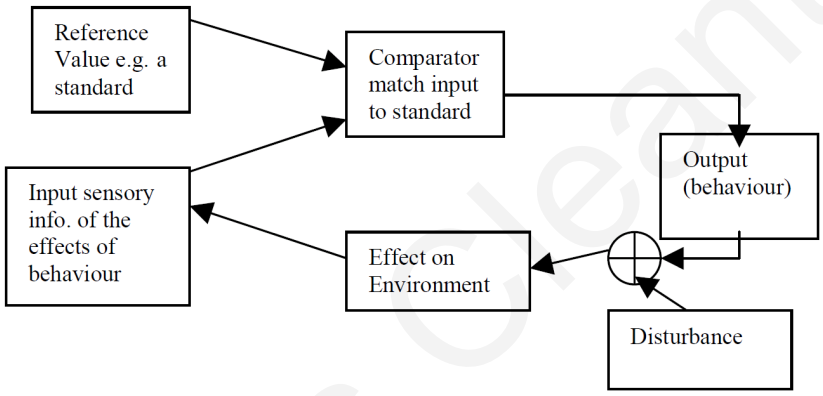

Figure 2.2: An abstract behavioural model of self-regulation
Output behaviour has certain effects on the environment. Input information about the effects of the behaviour on the environment is compared with a standard (moral or other) in order to determine whether the standard is maintained and change the behaviour (if needed) accordingly (adapted from Carver and Scheier, 1998).

From the viewpoint of cognitive neuroscience, a model of self-control can described as in is Figure 2.3 (Raclin, 2000). The model tries to integrate the neuronal mechanisms that underlie the behaviour and is summarised as follows. Information providing the current state of the environment comes into the cognitive system (arrow 1) located in the higher centre of the brain, which represents the frontal lobes associated with rational behaviour such as analytic thinking, planning and control. This information combines with signals from the lower brain, representing the limbic system (including

memory from the hippocampus) that is associated with emotion and action selection. This travels back down to the lower brain and finally results in behaviour (arrow 2), which is rewarded or punished by stimuli entering the lower brain (arrow 3).

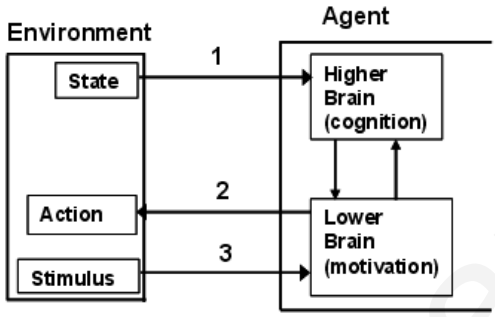

Figure 2.3: A model of self-control behaviour as an internal process, from the viewpoint of cognitive neuroscience

Information comes into the cognitive system (Arrow 1). This combines with the messages from the lower brain and memory and a choice is made, which results in behaviour (Arrow 2). The behaviour is finally reinforced (Arrow 3) (based upon Rachlin, 2000).

In the model, self-control behaviour is exhibited according to an interplay between the higher brain and the lower brain processes. Although this is consistent with the previously reviewed neuroimaging findings that identified both brain regions engaging in a self-control task, the model fails to provide any information on the dynamics, the qualitative and quantitative elements of the interplay that will enable self-control behaviour to emerge. The model is maybe too general and abstract that could also describe cases of decision making other than self-control.

The next model under consideration is the self-control strength model (Muraven and Baumeister, 2000). According to this model, the ability to exhibit self-control relies on a limited resource, or self-control strength, and all different self-control operations

draw on that same resource. In their previous study, Muraven et al. (1998) demonstrated that participants' performance was impaired in a self-control task that followed an initial one. In addition, the impairment was found even if the two tasks were completely different in context. The model's view of self-control resembles a muscle that its short-term ability decreases after exertion but at the same time repeated exercise strengthens it in the long-run. In another study (Muraven et al., 1999), a group of students was asked to regularly perform some easy self-control tasks for two weeks. These participants showed significant improvements on self-control compared with participants who did not practice self-control.

The model provides a powerful and useful analogy to self-control. The idea that self-control resembles a muscle that its exertion depletes a limited resource makes specific predictions with respect to self-control failure. In particular, people should tend to fail at self-control when recent demands and exertions have depleted their resource. Although Muraven et al. (1999) did not specify the nature of the limited resource on which self-control relies on, recent research (Gailliot and Baumeister, 2007) indicates that this resource is glucose. The study showed that reduced blood glucose and poor glucose tolerance (reduced ability to transport glucose to the brain) induced lower performances in self-control tasks.

The final model under consideration is provided by Kavka (1991). It is noted that despite the fact that it is an abstract, game theoretical model of internal conflict, it is thoroughly discussed in the remaining of this section as it encapsulates important aspects of internal conflict and is the one which inspired the implementation of our computational model.

Remember that in the beginning of this chapter we stated that the self can be perceived as a goal-directed hierarchical system, where goals are internally specified according to value systems (Scheier and Carver, 1988). In addition, the presence of more than one established value systems can divide the interest within a single individual and give rise to intra-personal conflict (Livnat and Pippenger, 2006). Self-control problems arise because human nature is not always rational as perceived in the context of economic theory which assumes rational utility agents (von Neumann and Morgenstern, 1947), otherwise the choice of the large delayed reward would have always been practiced. It is more appropriate to refer to human nature as multi-rational in the sense that the brain, as in the case of intra-personal conflict, can be viewed as a society of conflicting subagents (Minsky, 1985), each one of them selfishly seeking reward accumulation given its individual goal. In addition, it has been shown analytically, in a game theoretical context that an optimal brain can be composed of conflicting "selfish" agents (Livnat and Pippenger, 2006). Given all the above, a model that proposes that internal conflicts are resolved as if they were a result of strategic interaction among goal-directed subagents (Kavka, 1991), makes sense.

The model was used by Kavka (1991) to provide a psychological picture of how individuals who experience internal conflicts end up with suboptimal outcomes. In the current thesis work, we utilise this game theoretical view of internal conflict to investigate all the possible outcomes with respect to how a conflict is resolved with more emphasis on the optimal outcome which corresponds, as we will show below, to self-control behaviour.

For the purpose of an example of internal conflict consider the following case of conflicting value systems. A student faces a dilemma whether he or she should stay at

home and finish a project that is to be submitted the following morning or go to the pub and celebrate a friend's birthday.

Possible options can be:

a. Go to the pub and have fun.

b. Stay home and study.

c. Go to the pub for a quick drink and go back home and study.

d. Do nothing about it (preserve *Status quo*) or do something different. Other possible *d* outcomes could be staying at home but not be able to study or go to the pub and having a miserable time because of guilty feelings.

Now assume that these conflicting desires (or dimensions of evaluation) are represented by two distinct subagents; the academic-conscious agent and the fun-conscious agent. The student's academic-conscious agent orders these options according to their value as $b>c>d>a$, whereas the fun-conscious agent evaluates them as $a>c>d>b$. Although the underlying mechanisms performed in the student's real brain are highly complex and the knowledge about them is incomplete, the final decision of the student depends crucially on these assigned values. A realistic approach would be to suggest that such inner conflicts are resolved as if they were a result of strategic interaction among goal-directed subagents (Kavka, 1991). As Kavka (1991) stresses out, the validity of the model does not presuppose the existence of real subagents. The role of the internal agents is to help us understand how suboptimal outcomes with respect to internal conflicts are psychologically plausible. In other words, even if real subagents do not exist, treating internal conflicts as an interaction between internal rational agents will help us understand aspects of internal conflict such as the importance of the payoff structure on how the conflict is resolved.

According to Kavka (1991), in situations of intra-personal conflict, each of the subagents can either insist on getting their way or compromise to a choice that benefits the organism as a whole. According to our example, their interaction can be analysed and represented by theoretical games as in Table 2.1.

	Academic Agent		
Fun Agent		Compromise	Insist
	Compromise	c	b
	Insist	a	d

Table 2.1: Game theoretical representation of strategic interaction between subagents with conflicting value systems

Each agent can either "Compromise" (C) or "Insist" (I). There are four possible outcomes *a, b, c* and *d* or IC, CI, CC, II respectively, that result from the agents' combined choices.

The academic-conscious subagent can insist on staying at home and studying throughout the night or compromise to a choice involving less studying. On the other hand, the fun-conscious subagent can insist on partying throughout the night or compromise to a less fun outcome. If both agents decide to "Compromise" then the student goes to the pub for a quick drink and then goes back home and studies. This corresponds to the *c* outcome which is the second best outcome for each agent but the maximum for the individual as a whole. This outcome represents the case where the individual exhibited self-control. Although one would think that self-control would correspond to the *b* outcome, where the student stays home and studies, this would

maximise the payoff for the academic agent but not for the individual as a whole. In the case where any of the student's subagents decides to "Insist" in order to pursue its most preferred outcome (*a* for the fun agent and *b* for the academic agent), then there is the risk of ending up in the worse situation *d* if the other agent also decides to "Insist". Finally, if one decides to "Compromise" in order to achieve its second best outcome *c*, it also has to bear in mind that if the other agent chooses to "Insist" then the outcome will be the least preferred (*b* for the fun agent and *a* for the academic agent). The formal analysis of the game specifies that the *d* outcome is the only Nash equilibrium (Nash, 1950) of the game, so both agents will receive the inferior value of the *d* outcome whereas they could have achieved a superior in value *c* outcome, if they both "Compromised". Therefore, if the student's agents were faced by this dilemma for only one time and knew that this is the only time they would interact, they would have both "Insisted" and the student would preserve *Status Quo* or do something different. The possibility of obtaining an outcome other than the suboptimal outcome *d* lies in the fact that these agents are probably going to interact for an unspecified number of times.

Moreover, consider the payoff matrix of Table 2.2 where the four outcomes are replaced by the values that each subagent assigns to each outcome. The analysis remains the same; *a* and *b* are the most preferred outcomes for the fun and academic agent respectively, but *c* outcome is the best outcome for the individual as a whole (4+4 > 5+(-3)). Self-control behaviour is exercised by an individual in order to achieve a higher overall payoff which here would correspond to the *c* outcome where the individual gains 8 (4+4) compared to 2 (5+ (-3)), i.e., the total payoff for *a* or *b* outcome. In addition, these values are not absolute in the sense that a different set of subagents might apply different values to the same outcomes, thus the payoff matrix of Table 2.2 is just one of

an infinite number of possible matrices. However, the structure of the payoffs of any given matrix should preserve the agents' outcomes ordering.

	Academic Agent		
Fun Agent		Compromise	Insist
	Compromise	4, 4 (c)	-3, 5 (b)
	Insist	5, -3 (a)	-2, -2 (d)

Table 2.2: The payoff matrix of the interaction

Payoff for the fun agent is shown first. Notice that the outcome ordering for each agent is preserved as b (5) > c (4) > d (-2) > a (-3) for the academic agent and a (5) > c (4) > d (-2) > b (-3) for the fun agent.

This game theoretical representation of strategic interaction between subagents with conflicting value systems is based on a well studied theoretical game known as the Prisoner's Dilemma (PD) (Rappoport and Chammah, 1965), which has been used to model human cooperation (Axelrod and Hamilton, 1981) as well as intra-personal conflict (Kavka, 1991). In its standard one-shot version, the scenario of the PD unfolds as follows. Two people are arrested by the police under suspicion of a crime. They are kept into separate rooms where the investigator visits each one of them to offer the same deal: if one testifies for the prosecution against the other and the other remains silent, the betrayer goes free and the silent accomplice receives a major conviction. If both remain silent, both prisoners are sentenced for a minor charge. If each betrays the other, each receives a medium sentence. Each prisoner must make the choice of whether to betray the other or to remain silent. Both care much more about their personal freedom than about the

welfare of their accomplice. However, neither prisoner knows for sure what choice the other would make.

	Column Player		
Row Player		Cooperate	Defect
	Cooperate	R, R	S, T
	Defect	T, S	P, P

Table 2.3: Payoff matrix for the Prisoner's Dilemma

Payoff for the Row player is shown first. The game is defined by: Temptation to Defect (*T*) must be better than the Reward for Mutual Cooperation (*R*), which must be better than the Punishment for Mutual Defection (*P*), which must be better than the Sucker's payoff (*S*) (Rule: $T>R>P>S$) (see text for further description).

The PD is a game summarised by the payoff matrix of the Table 2.3. There are two players Row and Column. Each player has the choice of either "Cooperate"(*C*) (remain silent in the prison example) or "Defect" (*D*) (betray the other). For each pair of choices, the payoffs are displayed in the respective cell of the payoff matrix of Table 2.3. Payoff for the Row player is shown first. *R* is the "reward" payoff given when both cooperate. *P* is the "punishment" that each receives if both defect. *T* is the "temptation" that each receives if one by his/her own defects and *S* is the "sucker" payoff that one receives if he or she by his/her own cooperates. The only condition imposed to the payoffs is that they should be ordered such that $T>R>P>S$. Note that in general, game theory assumes rational players in the sense that each player wants to maximise his or her own payoff. In addition, each player knows the other is rational, knows that the other

knows he or she is rational, etc. In game theoretical terms, *DD* is the only Nash equilibrium outcome (Nash, 1950), whereas the cooperative *CC* outcome is the only outcome that satisfies Pareto optimality (Pareto, 1906). The "dilemma" faced by the players in any valid payoff structure is that, whatever the other does, each one of them is better off by defecting than cooperating. But the outcome obtained when both defect is worse for each one of them than the outcome they would have obtained if both had cooperated. In the latter interpretation, the "Compromise"-"Compromise" (CC) outcome corresponds to the self-control outcome which is the best for the organism (if we add up the two values) and the second best for each agent. Note also that since DD is the only Nash equilibrium outcome of the game, then CC can not be obtained given that the game is played only once. Apart from the payoff structure, the game specifies that one round of the game consists of the two players (agents) choosing their action simultaneously and independently and then informed about the outcome.

The Iterated Prisoner's Dilemma (IPD) is a game where the one-shot PD is played consecutively by two players. The design of the game requires an extra rule such that the cooperative outcome remains Pareto optimal. Namely, $2R>T+S$ guarantees that the players are not collectively better off by having each player alternate between "Cooperate" and "Defect".

For the purposes of the current work we model the infinitely iterated version of the game where the same game is repeated for an unspecified amount of rounds. The infinitely repeated version suits the interpretation of intra-personal conflict more realistically as the two internal agents compete with each other for more than one time and additionally they do not have any valid reason to believe that the next time they come into conflict would be the last time they will ever compete with each other. The formal

analysis of the infinitely iterated version shows that there are multiple equilibria including the CC (self-control) outcome (as opposed to the one-shot version where CC is unattainable), which now constitutes the best possible long-term outcome both for the organism and the agents individually. The latter is true because the possible outcome where one agent always "Defects" and the other "Cooperates" can never be sustained.

2.3 Related Work: Similarities and Differences with our Proposed Model

The proposed research builds on a recently awarded PhD (Banfield, 2006, supervised by Dr Christodoulou), where non-biologically realistic neural networks simulated self-control behaviour through competition in the IPD, as well as the effect of precommitment. The schematic model of Figure 2.3 was implemented as two feed forward multilayer perceptron type networks simulating two players, representing the higher and lower centres of the brain, competing against each other in the IPD game using reinforcement learning. It was a network architecture of two networks exhibiting different behaviours to represent the higher *versus* lower cognitive functions, as depicted in Figure 2.3. The higher brain centre (which is seen as far-sighted) is implemented with the Temporal Difference weight update rule (Sutton, 1998) with a lookup table whereas the lower brain centre (which is seen as myopic) is implemented with the Selective Bootstrap weight update rule (Widrow et al., 1973). The research made the theoretical premise that the higher and lower brain functions cooperate, i.e., work together, which is in contrast to the traditional view of the higher brain functioning as a controller overriding the lower brain. Given this model, precommitment behaviour can be viewed as resolving some internal conflict between the functions of the lower and the higher centres of the brain by restricting or denying future choices and hence can be thought of as resolving an internal

conflict by prevention. It does this by biasing future choices to the larger, but later reward. By applying a differential bias to the payoff matrix of the IPD, the precommitment effect was simulated in the computational model. The results showed that increasing the precommitment effect increases the probability of cooperating with oneself in the future (Christodoulou et al., 2009).

The current thesis shares some obvious similarities with Banfield's (2006) work but also has significant differences, particularly in the way we interpreted the problem of self-control in our computational models. The theme of our work is the same: modelling internal conflict in a functionally-decomposed computational system of two networks that competed in the IPD. In our work this same set-up is used to implement a different abstract model. The current system implements Kavka's (1991) model where the competition is between internal subagents of the brain, as opposed to Rachlin's (2000) model where competition is between the higher and lower centres of the brain. The difference is conceptually fundamental since we believe that a subagent of the brain makes use of both brain regions. In the work of Banfield (2006), when applying the example of the student (which we use as well, see Section 2.2), the hypothesis is made that each agent (fun and academic) corresponds to each brain region. In our work each agent corresponds to both regions where the action of cooperation corresponds to a stronger activation by the fronto-parietal regions and the action of defection corresponds to a stronger activation by the limbic areas. So in our model, as shown in Figure 2.4, the competition is between higher and lower processes of the brain within each subagent and also between subagents and not just between the two processes.

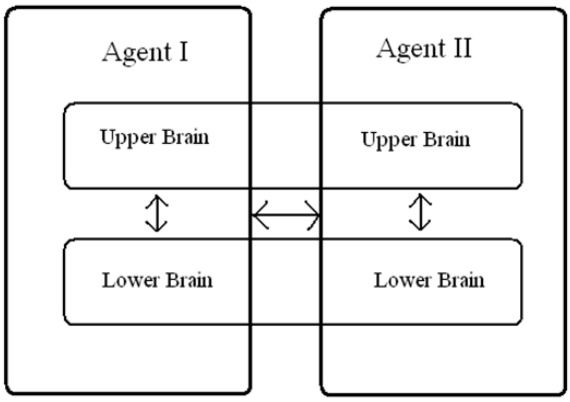

Figure 2.4: A schematic view of internal conflict as modelled by our computational model

Each agent can access brain areas concerning the valuation of both immediate and delayed rewards. Competition in the system (shown by arrows) exists between higher (delayed reward valuations) and lower processes (immediate reward valuations) of the brain within each subagent and also between subagents who serve distinct goals.

Another important difference is that in the student's example we use conceptually different options that correspond to different cells in the payoff matrix. This is not a problem of the particular example since it would apply to any example and displays a diverse notion with respect to self-control. Although the current thesis work and the work of Banfield (2006) agree that mutual cooperation (CC outcome) corresponds to the self-control outcome, in Banfield's (2006) example self-control corresponds to the student staying home and study. If we believe however that this is the best outcome for the academic agent why then the outcome where he stays at home but not able to study (when academic agent defects and the fun agent cooperates) has a greater payoff (correctly, as prescribed by the rules of the IPD)? Here lays a contradiction. Given our interpretation, self-control should not be the best outcome for any of the two subagents, but the best long

term outcome for the organism as a whole. This makes the payoffs in the payoff matrix of the game to be consistent with the outcomes in the student's example.

Moreover, in the current thesis we do not attempt to model self-control by precommitment while using the IPD since by definition precommitment is exercised in order to prevent a subsequent internal conflict (by restricting the available actions) whereas the IPD models the internal conflict itself. Therefore, in this thesis we investigate self-control behaviour only when experiencing an internal conflict.

A final difference is that although we implement the same multiagent reinforcement learning task (IPD), we use different kind of networks and learning algorithms. The competing agents in our work are implemented through biologically realistic spiking neural networks which learn through biologically plausible learning algorithms as opposed to the case of Banfield (2006), where classic artificial neural networks and standard machine learning algorithms are employed.

2.4 Reinforcement Learning on Spiking Neural Networks

In contrast to the case of traditional neural networks, it is only recently that reinforcement learning (RL) (Sutton and Barto, 1998) has been successfully applied to spiking neural networks. These schemes achieve learning by utilising various biological properties of neurons whether this is neurotransmitter release (Seung, 2003), spike timing (e.g. Florian, 2007) or firing irregularity (Xie and Seung, 2004). Their degree of experimental justification varies and it needs to be further assessed; nevertheless all these methods are biologically plausible and constitute the basis of successful RL application on biologically realistic neural models. A popular implementation of RL on spiking neural networks is achieved by modulating spike-timing-dependent synaptic plasticity (STDP)

with a reward signal (Faries and Fairhall, 2007; Florian, 2007; Izhikevich, 2007; Legenstein et al., 2008; Pfister et al., 2006). STDP is the change in synaptic efficacy which occurs according to the relative timing of pre- and postsynaptic spikes and has been experimentally observed. (Markram et al., 1997; Bi and Poo, 1998; Dan and Poo, 2004). Typically, STDP causes the potentiation of a synapse when the postsynaptic spike follows the presynaptic spike within a time window of the order of milliseconds, and the depression of the synapse when the order of the spikes is reversed (Hebbian STDP). It is also antisymmetric, because the sign of the modification reverses when the relative timing reverses. Experiments have also found synapses with anti-Hebbian STDP (opposite sign modifications in comparison to Hebbian STDP), as well as synapses with symmetric STDP (Dan and Poo, 1992; Bell et al., 1997).

Other examples of RL on spiking neural networks include Seung's reinforcement of stochastic synaptic transmission (2003) as well as reinforcement of irregular spiking (2004), where the learning rules perform stochastic gradient ascent on the expected reward by correlating the neurotransmitter release probability and the fluctuations in irregular spiking respectively with a reward signal. Vasilaki et al. (2009) uses a policy gradient method with a Hebbian bias whereas Potjans (2009) in a different study, a spiking neural network implements an actor-critic TD learning agent. These algorithms were shown to be able solve simple tasks like the XOR problem (Seung, 2003; Florian, 2007). In addition, reward-modulated STDP could learn arbitrary spike patterns (Faries and Fairhall, 2007) or precise spike patterns (Legenstein et al., 2008) as well as temporal pattern discrimination (Legenstein et al., 2008) and could be used in simple credit assignment tasks (Izhikevich, 2007). In the current thesis work we employ Seung's (2003)

reinforcement of stochastic synaptic transmission and Florian's (2007) reward-modulated STDP with eligibility trace.

To the best of our knowledge this is the first time that these algorithms are tested in a complex Multiagent Reinforcement Learning (MARL) task, as the IPD. In MARL the problem lies in the dynamic environment created by the presence of more than one learning agent. Such an environment is affected by the actions of all agents, thus, for a system to perform well, the agents need to base their decisions on a history of joint past actions and on how they wish to influence future ones. In MARL there could be different kinds of situations: fully competitive or adversarial (which could be modelled with zero-sum games), fully cooperative or coordinative (which could be modelled with team games), and a mixture of both (which could be modelled with general-sum games) such as the IPD. MARL has been an active and intense research area over the last years, during which numerous successful learning algorithms have been designed. Some examples of well known learning algorithms that however do not concern application in spiking neural networks include minimax-Q (Littman, 1994), Nash-Q (Hu and Wellman, 2003) and FoF-Q (Friend-or-Foe Q) (Littman, 2001).

Choosing the right spiking neuron model, when building a spiking neural network is extremely important (Izhikevich, 2004). In our case, given the complexity of our spiking neural network system, the LIF neuron model (Lapique, 1907; Stein, 1967) was chosen as the basic node of each spiking neural network, due to its simplicity and computational effectiveness compared to the more biologically detailed conductance-based models like the Hodgkin and Huxley model (1952) or even spiking neuron models of intermediate complexity like the Izhikevich model (2003) (used in a spiking network model by Arena et al.(2009)) or the model proposed by Christodoulou et al. (2002) or the

McGregor model (McGregor and Oliver, 1974; McGregor, 1987) (used in a network of spiking neurons by Lin et al. (1998) and by Swiercz et al. (2006)).

Chapter 3

Computational Model of Internal Conflict: System Design and Testing

3.1 Overview

In order to investigate the behaviour of self-control, we propose a novel computational model of internal conflict. The model integrates for the first time knowledge from such diverse areas such as psychology, game theory, neuroscience and computational neuroscience and is applied on a multiagent reinforcement learning task. Given related work in the respective areas, a computational model of interpersonal conflict is proposed where we implement two spiking neural networks as two players, learning simultaneously but independently, competing in the Iterated Prisoner's Dilemma (IPD) game. In this chapter we present the system's design with respect to its architecture and the IPD implementation. Moreover, we present the learning algorithms that are employed for the purposes of training the spiking neural networks as well as their testing for correct implementation. We finish by investigating our hypothesis that high irregularity at high firing rates enhances learning.

3.2 A Computational Model for Internal Conflict

The model simulates competition between internal agents of the brain. Kavka (1991) proposed that these agents compete in a certain game theoretical interaction, the IPD. The payoff structure of the game is the key for choosing this particular game to model internal conflict (for details see Chapter 2). In my opinion this is a simple, elegant, consistent and powerful way to model such a highly complex state of mind. The model captures the essential feature of self-control, given that the agents are required to postpone immediate gratification in order to attain delayed but more valuable outcomes. The choice of short term *versus* long term reward is represented in the structure of the payoff matrix of the IPD (Table 3.1). The action of defection yields the best possible immediate payoff in the case of unilateral defection by any of the two agents. Therefore each agent is tempted to Defect in order to collect the greater immediate payoff. However the case where one agent repeatedly Defects and the other Cooperates can never be sustained as the best response by the agent who unilaterally Cooperates is also to Defect in the succeeding rounds of the game such that to avoid the Sucker's payoff provided by unilateral cooperation. In such a case, the two agents will end up receiving the penalty of mutual defection. Therefore, none of the agents can attain long term reward maximisation through the action of defection because although it is possible to receive maximum immediate payoff through unilateral defection, it will trigger a behaviour of mutual defection which is far from an optimal one. In contrast, repeated mutual cooperation is stable. The agents repeatedly receive their second best immediate payoff, and refrain from defecting because they know that if they deviate to unilateral defection in order to gather the best immediate payoff then the other agent will also Defect in the succeeding rounds, as explained. Consequently, long term reward maximisation for each agent can only be

attained through mutual cooperation where the agents constantly receive their second best immediate payoff. Therefore, the action of cooperation corresponds to long term maximisation and the action of defection to short term maximisation. For these reasons, the choice of long term *versus* short term reward is implemented in our computational model through modelling the choice between the actions of cooperation and defection respectively.

These goal-driven agents (e.g., fun, knowledge, taste, fitness etc.) are rational, meaning that they pursue long term maximisation of their individual expected payoff. For the purposes of the current thesis the word "rational" is used in the economic context of utility theory, as specified in the previous sentence, and it is not used to mean reasoning, which implies syntax. Therefore, the implemented agents are rational, meaning that they 'selfishly' pursue reward maximization according to their own value systems that satisfy their individual goals, giving no interest on the well being of the other agent or the organism as a whole. Fortunately the well being of the organism is in line with their individual well being in the long run. This is because the agents can achieve long term payoff maximisation only through mutual cooperation which maximises payoff for the whole system as well. As Livnat and Pippenger (2006) analytically showed, an optimal brain can be composed of internal competing agents. Optimality in the case of internal conflict is the behaviour of self-control that can be attained if the agents follow a compromising strategy in their interaction through mutual cooperation.

Each of the internal agents or subagents of the brain is implemented in our computational system by a spiking neural network. Brain image studies identified two separate brain systems involved in decisions under internal conflict, one composed of limbic system areas and one of fronto-parietal ones (e.g. McClure et al., 2004). The

former is activated when decisions involve only immediate available rewards and the latter is activated irrespective of the delay of the rewards. Therefore it would be tempting to map each one of these spiking neural networks to each one of these neural systems. However, that would also mean that each agent would correspond to only one of these systems. Nevertheless, the limbic system has been shown to be activated only when considering immediate rewards whereas the agents in the IPD are required to consider both the immediate and the long term rewards. Therefore, even if it made sense to consider areas of the limbic system mapping to an agent, it would also make sense for that agent to respond only and always to the immediate reward and hence always to defect in the game. This would have catastrophic consequences since the best response to an "always Defect" strategy would be "always Defect" and would make the self-control outcome (mutual cooperation) unattainable. Hence, if we believe that self-control is possible in cases of internal conflict, and the IPD models this internal competition, such a mapping is defective.

Another reason for avoiding mapping each network to only one brain region becomes clear if we consider individuals like addicts who often engage reason in order to construct elaborate plans so as to acquire their dose. Even if drug taking can be regarded as an immediate gratification behaviour, the agent whose goal is drug consumption employs top-level processes alongside the low-level ones. Therefore we can not consider such an agent to be represented solely by the limbic system areas.

For all the above reasons, in our model (unlike the model of Banfield, 2006) each agent does not correspond to particular brain regions in a constricted sense. Decision making with respect to internal conflict involves the participation of many brain regions and the understanding of such a complex interaction necessitates further investigation.

However, experimental findings with respect to intertemporal choice (e.g. McClure et al., 2004) should not be disregarded. These findings are integrated in our model in the process of action selection. More specifically, for each network the action of cooperation corresponds to a stronger activation by the higher brain (since cooperation achieves delayed gratification) and the action of defection corresponds to a stronger activation by the lower brain (as defection yields immediate gratification). In addition, given that the decision of the subjects depends on the relative activation of the engaged systems, each network in our model decides whether to cooperate or defect according to the relative activation of its output units.

Therefore in our approach (unlike that of Banfield, 2006), the competition between the higher and lower processes of the brain has two dimensions; competition of cooperation and defection within each agent, but also across the agents where the final decision by the system depends on the overall competition between cooperation and defection. In our system this is implemented by the competition between the output units of each network whose relative activation is responsible for action selection and across the networks that compete in the IPD whose combined output activations determines the overall outcome. Therefore, although we implement Kavka's theoretical model of internal conflict where the competing agents have similar nature (as both agents implement both the higher and the lower processes of the brain through the actions of cooperation and defection respectively), the model does not contradict overall competition between higher and lower processes of the brain as for example in McClure's (2004) experimental work or in Rachlin's (2000) theoretical model.

The model's architecture is depicted in Figure 3.1. Each network has a multilayer perceptron type architecture with a hidden layer of 60 leaky Integrate-and-Fire (LIF)

neurons (Lapique, 1907; Stein, 1967) and an output layer consisting of 2 LIF neurons. The two networks share a common input layer of Poisson spike trains. Each network has full feed-forward connectivity between its three layers. The output layer is also the decision layer as the decision whether to cooperate or defect at a given round of the game is taken according to the relative activation of the two units. In general, the system's simple architecture shares very little with the complex structure of the brain, nevertheless in the course of this thesis we will try to demonstrate that it exhibits essential features of a highly complex behaviour of the human brain.

The networks learn simultaneously but separately where each network seeks to maximise its own accumulated reward. The game is simulated through an iterative procedure which starts with a decision by the artificial agents, continues by feeding this information to the agents, during which learning takes place, and ends by a new decision.

The agents take their first decision randomly. During each learning round, the input to the system is presented for 500ms and encodes the decisions the two networks had during the previous round. This means that after round *k*, the outcome of the game (at round *k*) is fed into the system for 500ms and the learning variables are changed accordingly. For example, if at a given round network *I* chooses to defect (*D*) and network *II* to cooperate (*C*), then during the next learning round the networks will receive input that encodes the defect-cooperate (*DC*) outcome.

The decision of each network is encoded in the input, by the firing rate of two groups of Poisson spike trains. The first group will fire at 40Hz if the network cooperated and at 0Hz otherwise. The second group will fire at 40Hz if the network defected and at 0Hz otherwise. Consequently, as shown in Figure 3.2, the total input to the networks

during each round is represented by four groups of Poisson neurons, two groups for each network, where each group fires at 40Hz or 0Hz accordingly.

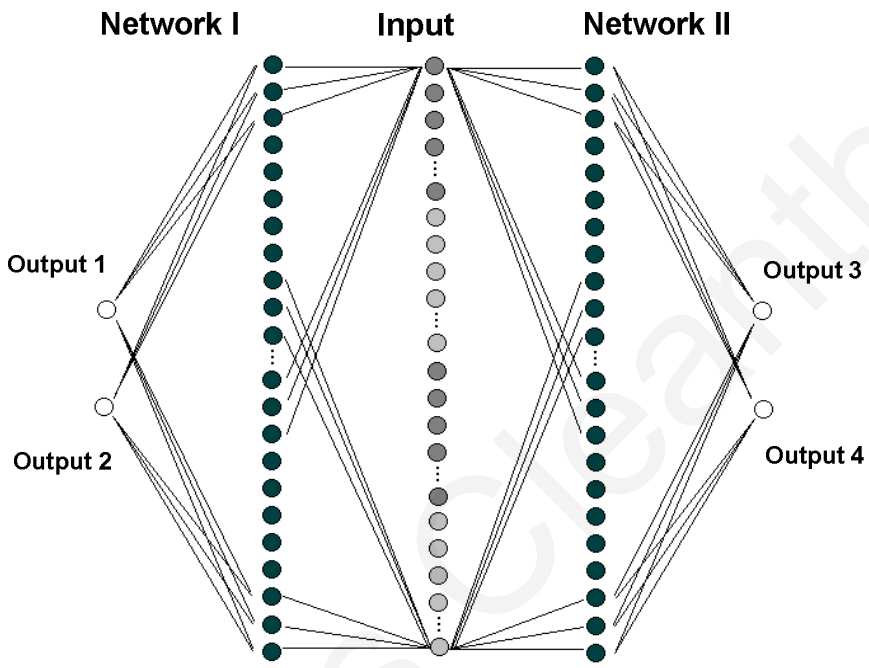

Figure 3.1: Computational model of internal conflict: representing two internal agents competing with each other

Two individual spiking neural networks with multilayer architecture receive a common input, depicted in the middle of the figure. Each network (left and right) has two layers that make feed forward connections between three layers of neurons; the 60 input neurons, 60 leaky integrate-and-fire hidden neurons and 2 leaky integrate-and-fire output neurons. The networks have full connectivity, though only some connections are shown for clarity. Neurons are randomly chosen to be either excitatory or inhibitory. The two networks simulate the conflicting agents. Actions for each agent are decided according to the relative activation of each network's output neurons.

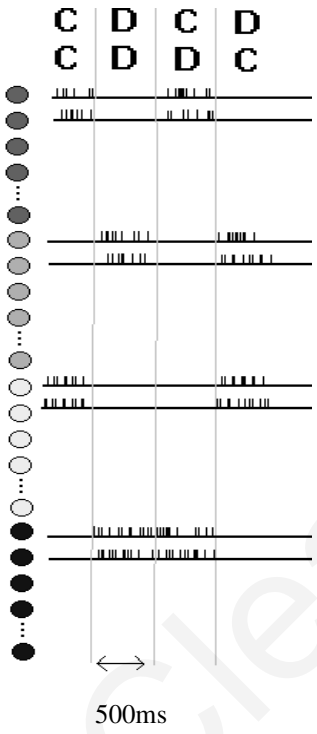

Figure 3.2: The input to the model

Four groups of Poisson spike trains encode the outcome of the game during the last round by firing at 40 or 0 HZ accordingly. For each network the first group will fire at 40Hz if the network cooperated and at 0Hz otherwise. The second group will fire at 40Hz if the network defected and at 0Hz otherwise. Each presentation lasts for 500ms.

For any given round there are always two groups of 40Hz Poisson spike trains, preserving thus a balance at the firing rates of the output neurons at the beginning of learning. Therefore, any significant difference in the firing rate of the output neurons at any time should be induced only by learning and not due to differences in the firing rates of the driving input.

At the end of each learning round the networks decide whether to cooperate or defect for the next round of the game. Decisions are carried out according to the value

that each agent assigns to the two actions, and these values are reflected in the firing rates of the output neurons. The value of cooperation for networks *I* and *II* is taken to be proportional to the firing rate of output neurons *1* and *3* respectively. Similarly, the value of defection for network *I* and *II* is taken to be proportional to the firing rate of output neurons *2* and *4* respectively. At the end of each learning round the firing rates of the competing output neurons are compared, for each network separately, and the decisions are drawn.

As stated, decisions for each agent are drawn according to the relative aggregate activation of its output units where the one unit responds to future rewards (since it evaluates the action of cooperation) and the other to the immediate rewards (since it evaluates the action of defection). The respective set-up is chosen so that it is consistent with findings in intertemporal choice experiments (McClure et al., 2004; Tanaka et al., 2004) where action was drawn according to the relative activation of two competing systems; one comprising fronto-parietal areas which is involved in evaluating future options and parts of the limbic system which are preferentially activated by decisions involving immediately available rewards. Thus we could think of the firing rates of the two output neurons as a reflection of the activation of these neural systems.

When the two networks decide their play for the next round of the IPD, they each receive a distinct payoff given their actions and according to the payoff matrix of the game (Table 3.1). The payoff each network receives as a result of their combined actions at the game also serves as the global reinforcement signal (scaled down) that will train the networks during the next learning round and thus guide the networks to their next decisions. The payoffs are scaled down when administered as reinforcements to the networks in order to incorporate the distinction between signals given by the environment

and how these signals are internally processed. The scaled down payoffs combined with a small learning rate ensure that changes on the variables controlled by learning are made in a smooth and gradual way.

	Agent II		
		Cooperate	Defect
Agent I	Cooperate	4, 4	-3, 5
	Defect	5, -3	-2, -2

Table 3.1: Payoff matrix of the interaction

The agents are competing in the IPD according to this particular payoff structure. Payoff for the Agent *I* is shown first. Mutual cooperation is the best outcome for the system as a whole whereas unilateral defection brings the best outcome for the defector.

For example, if the outcome of the agents was a CD, then according to the payoff matrix network I should receive a payoff of -3 for cooperating and network II a payoff of +5 for defecting. As stated, the reinforcement signal is specified according to the aggregate activation of the output units at the end of a learning round since the decision of the agents whether to cooperate or defect depends on the aggregate relative activation of each network's output units. This reinforcement is constant in value during the next 500ms of learning and is applied in the timestep following the spikes of the output neurons, as prescribed by the original learning algorithms (Seung, 2003; Florian, 2007). In addition, each network is reinforced for every spike of their output neuron that was "responsible" for the decision at the last round and therefore for the payoff received.

Hence in the CD case, network I would receive a constant penalty of -3 (scaled down to -1.3) that is applied for every spike of output neuron 1 (remember that the firing rate of output neuron 1 reflects the value that network I has for the action of cooperation) and network II would receive a constant reward of +5 (scaled down to 1.5) applied for every spike of output neuron 4 (remember that the firing rate of output neuron 4 reflects the value that network II has for the action of defection).

Since the learning algorithms work with positive and negative reinforcements, it is necessary that the payoff matrix contains both positive and negative values. The networks therefore learn through global reinforcement signals which strengthen the value of an action that elicited a reward and weaken the value of an action that results in a penalty.

It should be stressed out that the training of the system is not performed in the conventional way when one assumes reinforcement learning. A typical procedure would be to present the input for a given amount of time and observe the output, then compute the payoff, present the same input again and apply learning according to the computed payoff. In our case, we present the input for a given amount of time (which encodes the agents' last decisions) and at the same time reinforce the system according to the payoff which is computed according to the agents' last decisions. We believe that despite that our system learns in a non conventional way, our training scheme is more appropriate in the case of a game theoretical interaction such as the IPD. With our training scheme, learning is always on-line and the agents directly learn whether to cooperate or defect as a best response to the other agent's action rather than associate specific inputs to specific outputs. For example, if the outcome of a given round is CD then reinforcement is administered such that the action of defection for both agents is promoted (since according to the payoff matrix of the game, agent 1 receives maximum penalty for

cooperating and agent 2 receives maximum reward for defecting). Therefore, learning will induce a DD output which is not optimal but it makes sense in game-theoretical terms as it is the best response of both agents to the other agent's action. If training was performed according to the typical way, then in the case of having CD as an input with no learning taking place, a possible output could also be a CC for which both agents should receive positive rewards (according to the payoff matrix of the game) when learning is subsequently applied. Consequently, learning could associate a CC output for the CD input which is the optimum and preferred outcome for the system in terms of accumulated reward, but it is unrealistic in game theoretical terms, when considering best responses by the agents. Our aim is to induce a strong CC behaviour between the agents but in a way that is realistic and consistent with game theory and not just maximise the reward of the system.

Figure 3.3 shows how training is performed during two learning rounds. Suppose that the networks' last decision was a DC. The input to the system encodes this decision by four groups of Poisson spike trains firing either at 40 Hz or 0 Hz. For a DC input the second and third group fire at 40 Hz. This input is presented for 500ms with simultaneous learning taking place where constant reinforcement is administered according to the DC previous outcome. Reinforcement is applied in the timestep following every output spike of the neurons responsible for that decision which in this case are output neurons 2 and 3. At the end of the 500ms the aggregate activation of the output neurons for each network is compared and the new decision of the networks is drawn. In our example, output neuron 2 and 4 had a greater total activation therefore the networks' new decision is DD. Therefore, the new input to the system is the one encoding the DD outcome where the second and fourth group fire at 40 Hz. Constant reinforcement is administered for 500ms

according to the DD outcome and therefore applied in the timestep following spikes of output neurons 2 and 4. At the end of the 500ms the new outcome is drawn which in our example is a CC since the aggregate activation of output neurons 1 and 3 were greater.

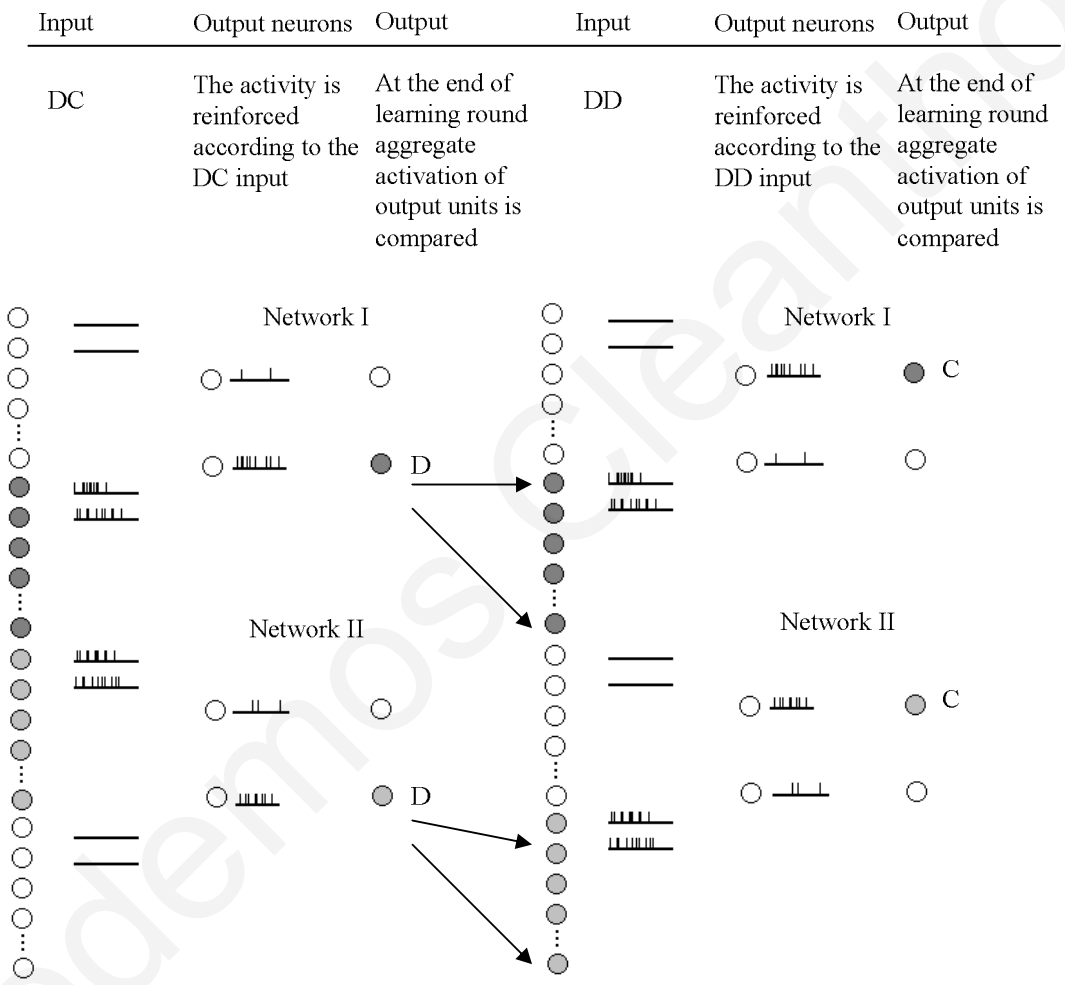

Figure 3.3: Sample training procedure

The figure presents sample training for two learning rounds. Four groups of Poisson spike trains feed the system with input that encodes the networks' last decision (active groups are shown in gray). Shown next are the four output neurons and their activity, two for each network. At the end of the learning round decisions are drawn where greater activation for each network is shown in gray. Arrows show how the outcome of the learning round is translated into the next input.

3.3 Learning Algorithms Employed for Training the Agents

Reward maximization for each agent is enabled by learning. The current work employs reward-modulated STDP with eligibility trace (Florian, 2007) and reinforcement of stochastic synaptic transmission (Seung, 2003). Both algorithms are derived as an application of the online partially observable Markov decision process (OLPOMDP) reinforcement learning (Baxter et al., 2001) algorithm and also keep a record of the agents' recent actions through the eligibility trace. In reward-modulated STDP the agent is regarded to be the neuron that acts by spiking and the parameter that is optimised is the synaptic connection strength. On the other hand, in reinforcement of stochastic transmission the synaptic connection strengths are constant, the agent is regarded to be the synapse itself that acts by releasing a neurotransmitter vesicle and the parameter that is optimised is one that regulates the release of the vesicle. To the best of our knowledge this is the first time that these algorithms are applied to a demanding MARL task.

3.3.1 Reward-modulated STDP with Eligibility Trace

In reward-modulated STDP with eligibility trace (Florian, 2007) the modulation of standard antisymmetric STDP with a reward signal leads to RL. The synaptic efficacies exhibit Hebbian STDP when the network is rewarded and anti-Hebbian when punished, allowing the network to associate an output to a given input only when accompanied by a positive reward and disassociate one when accompanied by a punishment, permitting thus the exploration of better strategies. Moreover it involves a biological plausible variable, the eligibility trace (Klopf, 1982) that serves as a decaying memory of the correlations between recent pre- and postsynaptic spike pairs.

According to Florian (2007) in reward-modulated STDP with eligibility trace, the efficacy of the synapse from neuron *j* to *i* is changed according to equation 3.1:

$$w_{ij}(t+\delta t) = w_{ij}(t) + \gamma\ \delta t\ r(t+\delta t)\ z_{ij}(t+\delta t) \quad (3.1)$$

where γ is the learning rate, δt is the duration of a time step, r is the global reward signal and z is the eligibility trace which is modified according to equation 3.2:

$$z_{ij}(t+\delta t) = \beta z_{ij}(t) + \zeta_{ij}(t) / \tau_z \quad (3.2)$$

β is a discount factor between 0 and 1, ζ is a notation for the change of z resulting from the activity in the last time step and τ_z is the time constant for the exponential decay of z. At time t, ζ is computed by the following set of equations (3.3-3.5) where the variable P^+_{ij} tracks the influences of presynaptic spikes and the variable P^-_{ij} tracks the influence of postsynaptic spikes. The time constants τ_+ and τ_- determine the ranges of interspike intervals over which synaptic changes occur and according to the standard antisymmetric STDP model, A_+ and A_- are positive and negative constant parameters respectively. Finally $f_i(t)$ is 1 if neuron *i* has fired at time step t or 0 otherwise.

$$\zeta_{ij}(t) = P^+_{ij} f_i(t) + P^-_{ij} f_j(t) \quad (3.3)$$

$$P^+_{ij}(t) = P^+_{ij}(t-\delta t)\exp(-\delta t / \tau_+) + \mathrm{A}_+ f_j(t) \quad (3.4)$$

$$P^-_{ij}(t) = P^-_{ij}(t-\delta t)\exp(-\delta t / \tau_-) + \mathrm{A}_- f_i(t) \quad (3.5)$$

The networks are composed of integrate-and-fire neurons with resting potential u_r = -70 mV, firing threshold θ = -54 mV, reset potential equal to the resting potential and decay time constant τ = 20 ms. These are the same values as used in the simulations by Florian (2007). We also used the same dynamics for the neurons' membrane potential given by equation 3.6:

$$u_i(t) = u_r + [u_i(t-\delta t) - u_r]\exp(-\delta t/\tau) + \sum_j w_{ij} f_j(t-\delta t) \quad (3.6)$$

The membrane potential was reset to u_r when surpassed θ. We used $\tau_+ = \tau_- = 20$ ms, $A_+=1$ and $A_- =-1$, $\delta t = 1$ ms, $\gamma = 0.7 \times 10^{-4}$ and unless specified, $\tau_z = 25$ ms.

3.3.2 Reinforcement of Stochastic Synaptic Transmission

In reinforcement of stochastic synaptic transmission, Seung (2003) makes the hypothesis that microscopic randomness is harnessed by the brain for the purposes of learning. The model of the hedonistic synapse is developed along this hypothesis. Briefly, within the framework of the model, each synapse acts as an agent who pursues reward maximisation through the actions of releasing or not a neurotransmitter. Synapses effectively learn by computing a stochastic approximation to the gradient of average reward. Moreover, if each synapse behaves hedonistically then the network as a whole behaves hedonistically, pursuing reward maximisation.

Upon arrival of a presynaptic spike, a synapse can take two possible actions with complementary probabilities; release a neurotransmitter with probability p or fail to release with probability 1 - p. The release parameter q is monotonically related to p by the sigmoidal function given by equation 3.7:

$$p = \frac{1}{1+e^{-q}} \quad (3.7)$$

Each synapse keeps a record of its recent actions through a dynamical variable, the eligibility trace (Klopf, 1982) which signifies when a synapse is eligible for reinforcement by keeping a record of the synapse's recent actions with respect to neurotransmitter release. It increases by 1- p with every release and decreases by $-p$ with

every failure. Otherwise it decays exponentially with a given time constant. When a global reinforcement signal is given to the network, it is subsequently communicated to each synapse which modifies its release probability according to the nature of the signal (reward or penalty) and its recent releases and failures. Learning is driven by modifying q according to the rule given by equation 3.8:

$$\Delta q = \eta \times h \times \bar{e} \quad (3.8)$$

where η is the learning rate, h is the reinforcement signal and $\bar{e}$ the eligibility trace.

Each network has a hidden layer of 60 neurons and an output layer of 2 neurons, all modelled with the leaky integrate-and-fire equation (3.9):

$$C\frac{dV_i}{dt} = -g_L(V_i - V_L) - \sum_j G_{ij}(V_i - E_{ij}) \quad (3.9)$$

where V_L= -74 mV, g_L= 25 nS and C= 500 pF. The differential equations are integrated using an exponential Euler update with a 0.5 ms time step. When the membrane potential V_i reaches the threshold value of -54 mV, it is reset to -60mV (values as in the numerical simulations by Seung, 2003). The reversal potential E_{ij} of the synapse from neuron j to neuron i is set to either 0 or -70 mV, depending on whether the synapse is excitatory or inhibitory. The synaptic conductances are updated via $\Delta G_{ij} = W_{ij}\, r_{ij}$ where r_{ij} is the neurotransmitter release variable that takes the value of 1 with probability equal to the probability that the synapse from neuron j to i releases a neurotransmitter (when j spikes) and 0 otherwise (Seung, 2003). In the absence of presynaptic spikes G_{ij} decays exponentially with time constant τ_s = 5 ms. W_{ij} are the "weights" which do not change over time and are chosen randomly from an exponential distribution with mean 14nS for excitatory synapses and 45nS for inhibitory synapses.

3.4 Testing the Learning Algorithms for Correct Implementation

The learning algorithms were tested for correct implementation in the classic benchmark problem of XOR. The XOR function performs the following mapping between two binary inputs and one binary output: {0, 0}→ 0; {0, 1}→ 1; {1, 0}→ 1; {1, 1}→ 0. The network architecture for testing the two algorithms was the same as the one Seung (2003) and Florian (2007) used for the same problem i.e, a feedforward neural network with 60 input neurons, 60 hidden neurons and one output neuron. Each layer had full feed-forward connectivity to the next one. Neurons were randomly selected to be either excitatory or inhibitory. The first 30 input neurons encoded the first binary input and the rest the second input. The input "1" was encoded by a Poisson spike trains firing at 40 Hz, while the input "0" was represented by the absence of spiking. Each input presentation ({0, 0}, {0, 1}, {1, 0}, {1, 1}) lasted 500 ms.

The training was accomplished by presenting the inputs and then delivering reward or punishment to the synapses, according to the activity of the output neuron. More specifically the network was rewarded for every output spike when the input was either {0, 1} or {1, 0} and punished for every output spike when the input was {0, 0} or {1, 1}. In other words, training promoted output activation when desired and suppressed it otherwise. As in Florian (2007), we considered that the networks are able to solve the XOR problem, if at the end of an experiment, the output firing rate for the input pattern {1, 1} was lower than the output firing rates for the patterns {0, 1} or {1, 0}. The output firing rate for the input pattern {0, 0} was always 0, as a result of the rate coding of the input patterns (as in Florian, 2007). Before learning took place, the network naturally responded with more output spikes to input {1, 1} than to {0, 1} or {1, 0} since all 60 input neurons were firing at 40 Hz.

As shown in Figure 3.4 the network managed to compute the XOR function with reward- modulated STDP with eligibility trace (Florian, 2007) as well as with reinforcement of stochastic synaptic transmission (Seung, 2003). The synapses changed during (synaptic weight and release probability respectively) learning in such a way as to increase the reward received by the network, by suppressing or enhancing the output activation for the input patterns accordingly.

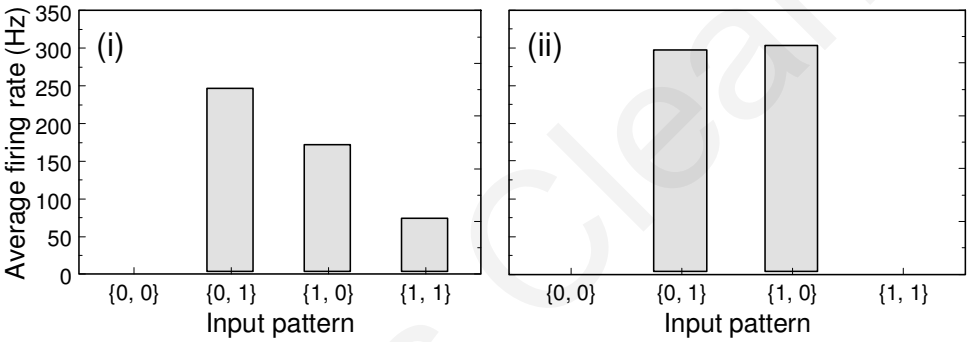

Figure 3.4: Learning the XOR computation

Average firing rate of the output neuron after learning, for the four different XOR input patterns. Both algorithms managed to efficiently compute the function although reinforcement of synaptic transmission performed better. (i) The first chart corresponds to learning with reward modulated STDP with eligibility trace. The learning rate is set to 0.00007 and the eligibility trace time constant τ_z to 25ms. (ii) The second chart shows the performance of reinforcement of stochastic synaptic transmission. The learning rate is set to 0.3 and the eligibility trace time constant τ_z to 20ms.

3.5 Investigating the Performance of Reward-Modulated STDP with Eligibility Trace: Does High Firing Irregularity Enhance Learning?

In the course of investigating the performance of modulated STDP with eligibility trace we explored the hypothesis that high firing irregularity at high rates would enhance learning. We believed that this is possible as high firing irregularity will lead to more accurate correlations between pre-synaptic and postsynaptic spike timings and reinforcement signals. If firing is regular, then it is possible for two identical spike pairs to be associated with opposite in sign reinforcement signals, confusing thus the direction of the plasticity for a given synapse. High firing irregularity prevents this unnecessary competition by weakening this possibility and thus preventing a possible corruption of the learning algorithm.

In order to better understand how regularity may destroy learning, we should observe the how the dynamics of the variables used by reward-modulated STDP with eligibility trace affect the synaptic strength, as presented in Figure 3.5. f_j shows a presynaptic regular spike train, f_i shows a postsynaptic regular spike train, the variable P^+_{ij} tracks the influences of presynaptic spikes and the variable P^-_{ij} tracks the influence of postsynaptic spikes. In addition z_{ij} is the eligibility trace, ζ_{ij} is a notation for the change of z_{ij} and finally w_{ij} is the synaptic strength. For more details regarding how the variables are computed and associated with each other please see Section 3.3.1 or the original paper (Florian, 2007). Figure 3.5 shows how the synaptic strength changes with time for two regular presynaptic and postsynaptic spike trains. The problem of such a case is evident if we observe the synaptic strength which oscillates around a given value until the sign of the reinforcement changes when it continues to oscillate around another value. Therefore, for the time period where a constant reward or penalty is administered, the effect of any

pre-post spike pair is cancelled out by the next one and the value of the synaptic strength remains effectively constant, destroying thus learning during that period of time.

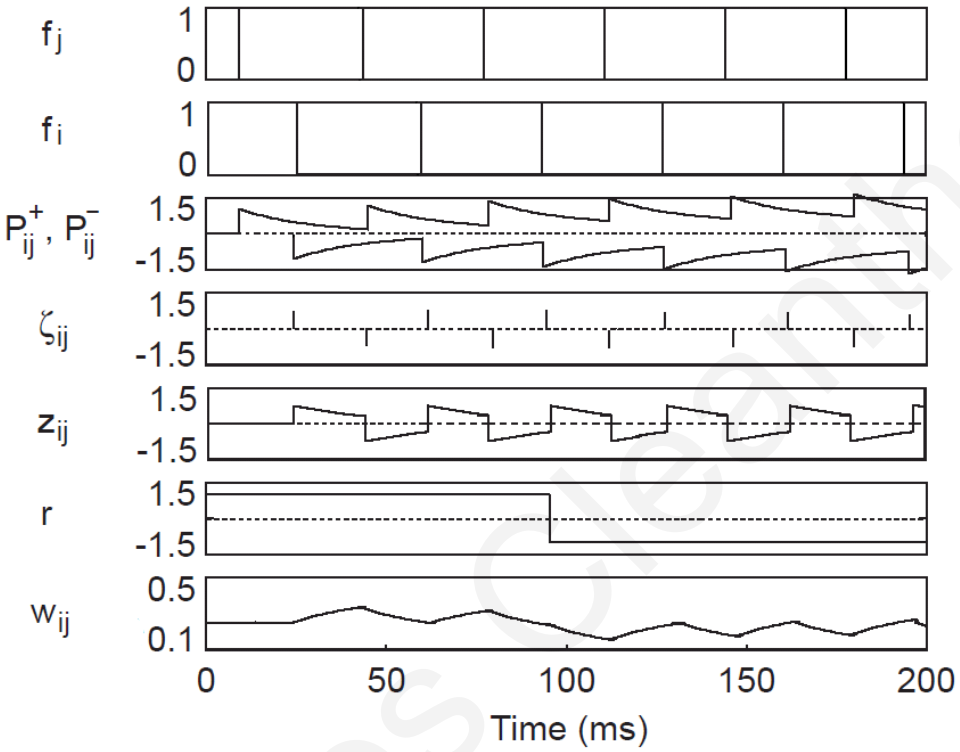

Figure 3.5: Effect of regularity in the value of the synaptic strength

An illustration of the dynamics of the variables used by reward-modulated STDP with eligibility trace and the effects on the synaptic strength when spike trains are regular. f_j shows a presynaptic spike train, f_i shows a postsynaptic spike train, P^+_{ij} tracks the influences of presynaptic spikes, P^-_{ij} tracks the influence of postsynaptic spikes, z_{ij} is the eligibility trace, ζ_{ij} is a notation for the change of z_{ij} and w_{ij} is the synaptic strength. Ths figure is a modified version of the one presented in the original paper for the learning algorithm (Florian, 2007) and is modified by us in order to explain how regularity can degrade learning.

In addition, if we consider the whole period of learning we observe another oscillation since the average change induced in the synaptic strength by the reward is later

cancelled out by the penalty and the value of the average synaptic strength is equal to its starting value, meaning that no learning took place. Here we presented a case where the magnitude of the reward equals the magnitude of penalty; this causes the synaptic strength to oscillate around its starting value. However, even in the case where the reward was unequal to penalty the synaptic strength would still oscillate around a given value, different from its starting one this time, but again learning would be degraded. Overall, regularity impairs learning because it makes the value of the synaptic strength to oscillate.

We achieved the required high firing irregularity of the LIF neuron by employing the partial reset mechanism. It has been shown (Bugmann, Christodoulou and Taylor, 1997; Christodoulou and Bugmann, 2001) that a LIF neuron model with partial somatic reset is a very good candidate for reproducing the observed highly irregular firing at high rates by cortical neurons (Softky and Koch, 1992, 1993).

In the current simulations, a high output firing rate of approximately 100Hz was targeted and achieved for both systems (with or without the partial somatic reset mechanism in their LIF neurons), which is within the high rate bound in which cortical cells *in vivo* fire irregularly as identified by Softky and Koch (1992, 1993). This was done by providing greater input frequency to the system comprising of LIF neurons with total reset, in order to compensate for the increased output firing rate in the other system due to the partial reset in its LIF neurons. Note that no direct comparison can be made between this system of total somatic reset and the one in section 3.4 (Figure 3.4 (i)) because of the different input frequencies.

The partial somatic reset mechanism works as follows: when the membrane potential $u(t)$ surpasses the firing threshold θ, then instead of being reset to the resting potential u_{rest}, it is reset to a level $u(t) = u_{rest} + \beta(\theta - u_{rest})$, where β is the reset parameter,

with a value between 0 and 1. For the purposes of our study we used $\beta = 0.91$; this value of the reset parameter was chosen as it was found to produce the observed high firing irregularity at high rates by cortical neurons (Bugmann, Christodoulou & Taylor, 1997; Christodoulou & Bugmann, 2001). More specifically, in Christodoulou & Bugmann (2001), it was showed that with the somatic reset value set at $\beta = 0.91$, the firing interspike intervals (ISIs) at high rates are: (i) exponentially distributed and (ii) independent; in addition, in Bugmann, Christodoulou & Taylor (1997), it was demonstrated that the coefficient of variation (CV) *vs* mean firing ISI curve with $\beta = 0.91$ shows a close similarity, firstly with the experimental one (Softky and Koch, 1992, 1993) and secondly with the theoretical curve for a random spike train with discrete time steps and a refractory time. In the respective simulations in this thesis the CV was approximately equal to 0.85. Therefore, with the choice of the reset parameter β set to 0.91, the firing ISIs are purely temporally irregular (and there are no bursts, that could increase the firing variability), which fulfills our aim to investigate whether high firing irregularity enhances learning. Thus $\beta = 0.91$ is the optimal reset value parameter for our purpose and there is no need to see the performance for other reset value parameters, apart of course for $\beta = 0$.

As it can be seen by the results for the XOR problem (Figure 3.6), even though both types of network learned the XOR function, the network with the partial somatic reset mechanism in its LIF neurons performed much better in the task, than the one comprising of LIF neurons with total reset. In particular, the former type of network displayed more qualitative results than the latter, as it managed to consistently suppress more the output firing rate for input pattern {1, 1}, leading to a bigger difference between the output firing rates for input pattern {1, 1} and input patterns {0, 1} or {1, 0}. More

specifically, in the network consisting of LIF neurons equipped with partial reset, the suppression of the output firing rate for input pattern {1, 1} reached 63% of the average output firing rates for input patterns {0, 1} and {1, 0}, while the respective suppression percentage of the network having LIF neurons with total reset reached only 10%.

Results show that when LIF neurons fire at high rates then the performance of a spiking neural network in computing the XOR function increases when the partial somatic reset mechanism is used in conjunction with the modulated STDP algorithm (Florian, 2007). This is due to the high irregular firing of the LIF neurons that enabled the algorithm to perform more accurate correlations between pre-synaptic and postsynaptic spike timings and reinforcement signals.

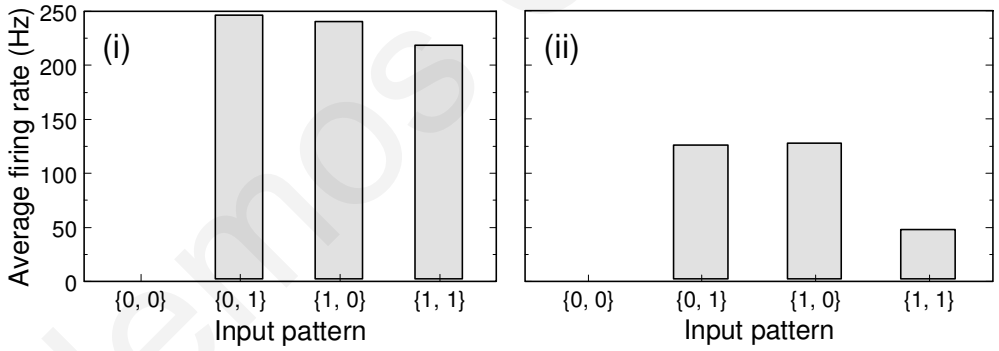

Figure 3.6: Effect of increased firing irregularity on learning the XOR computation Average firing rate of the output neuron after learning, for the four different XOR input patterns with the LIF neurons of the network having either total somatic reset (i), or partial reset with β = 0.91 (ii). In both problems the networks learn with reward-modulated STDP with eligibility trace (Florian, 2007), whose time constant, τ_z is set to 25ms for all networks and the learning rate to 0.00007.

Chapter 4

Simulating Internal Conflict

4.1 Overview

In Chapter 3 we described how we developed a computational model of internal conflict that implements the view (Kavka, 1991) that internal conflict can be resolved as if it was a result of strategic interaction between rational agents. The setting for this interaction is a well studied game, the IPD. We begin by describing how we overcame initial problems and continue by presenting results that demonstrate the ability of the system to exhibit self-control behaviour. Results are presented separately for the two implemented algorithms; first we present the results when learning was implemented by reinforcement of stochastic synaptic transmission (Seung, 2003) and continue by showing results obtained by reward-modulated STDP with eligibility trace (Florian, 2007). Moreover, we explore whether high firing irregularity at high rates in conjunction with the latter algorithm can improve the results of the system in this complex MARL task, as it did for the simple XOR problem in the previous chapter.

4.2 Reinforcement of Stochastic Synaptic Transmission

For the system configuration described in Chapter 3 a single game of the IPD consists of 200 rounds during which the two networks seek to maximise their individual accumulated payoff by cooperating or defecting at every round of the game. The following simulations involve implementation of the game where the agents learn through reinforcement of stochastic synaptic transmission (Seung, 2003). The learning rate used in the following simulations is 0.1.

It should be noted that the CC outcome is the best immediate outcome for the system as a whole, and it also maximises long-term reward for both the system and the agents individually. In addition, for the purposes of this thesis the consistent and persistent choice of the CC outcome during the IPD, specifies whether the agents exercised self-control behaviour. When we applied the algorithm as described in Chapter 3, in such a way as to reinforce only the actions that elicited a given outcome (for example if the outcome was CD, the action of cooperation for network *I* and the action of defection for network *II* were reinforced during the following learning round), the agents did not show the ability required to learn how to cooperate. The respective simulation results are shown in Figure 4.1.

The accumulated payoff is calculated by adding together the payoff each agent received according to the payoff matrix represented in Table 3.1 (Chapter 3). For example if at a given round the outcome was CC, then a total 4+4=8 will be added on the accumulated payoff. For the DC and CD outcome the total added payoff is 2 and for DD is -4. Given this, the system could achieve a maximum of 1600 (200 rounds × 8) if the two networks cooperated all the time. Results show that the system accumulated a total reward of less than 550 because of a low cooperative result (shown in Figure 4.2). The

CC outcome occurred only 31.5% of the time, which is a bit more than if it had occurred by chance (25%). This is because the agents did not learn how to cooperate in order to maximise their long-term reward and the system performed sub-optimally.

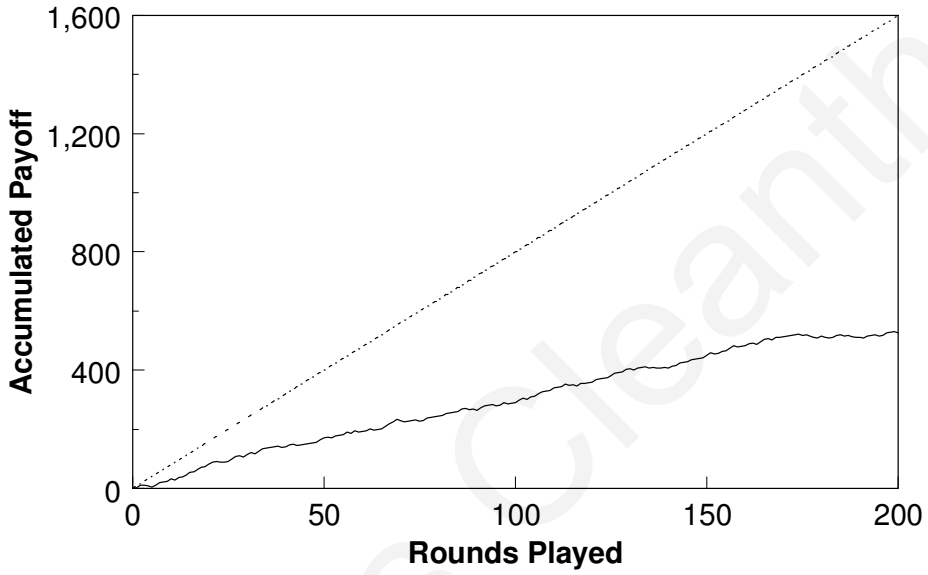

Figure 4.1: Total accumulated payoff, gained by both networks during the IPD

The system failed to accumulate a high payoff as it gained less than 550 out of a maximum 1600 (solid line). The agents did not learn to cooperate. The theoretically best performance is shown for comparison (*dot-dashed line*).

A closer examination revealed that at the end of each learning round both output neurons of each network exhibited approximately the same firing rate. This effect was due to lack of competition in the decision layer. Remember that training during a learning round aimed at modifying the activation of just the output neuron that was responsible for the decision. However, a single global reinforcement induced a parallel and similar alteration in the activity of the output neuron which was not intended to be altered during the learning round. This was happening because the activity of the neurons in the hidden

layer could not be prevented for changing the activity of both output neurons as (i) these hidden neurons feed to both output neurons and (ii) the subset of the synapses from the hidden layer to the output neuron whose activity was not intended to change were held fixed during the learning round, making them unable to neutralise the effect of the altered activation in the hidden layer. Therefore, any changes in the activity of the hidden layer due to learning, were propagated in the activity of both output neurons. The problem was tackled by enhancing the contrast between the activation of the output neurons through introducing additional global reinforcement signals that were administered alongside the original. These signals were also constant during the 500ms and were applied to the networks in the timestep following every spike of the output neurons that were not "responsible" for the decision at the last round. Overall during a learning round, each network receives global, constant and opposite in sign reinforcements that are applied for spikes of both of its output neurons. These two opposite in sign signals effectively push the activation of the output neurons in opposite directions, enhancing thus the contrast in their firing. The activation of the units in the hidden layer is now changed according to the activation of both of the output units and in addition, all the synapses from the hidden layer to both output neurons are now modified such that to control the activation of both output neurons in order to maximise received reward and minimise received penalty. For example in the *CD* outcome, an additional constant reward of +1.15 is applied to network *I* for every spike of output neuron *2* and an additional constant penalty of -1.15 is applied to network *II* for every spike of output neuron *3* (see figure 3.1 for output neuron numbering). The value of 1.15 applies to all outcomes and is chosen to be small and equal for all outcomes such that: (i) any changes to the values of the agents' actions (reflected

in the activation of the output neurons) are primarily induced by the reinforcement signals provided by the payoff matrix and (ii) the IPD payoff rules are not violated.

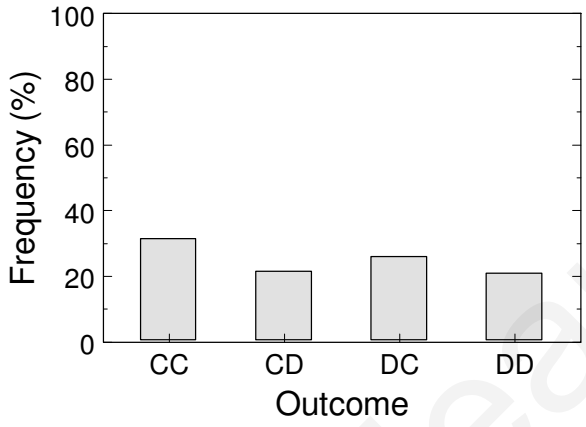

Figure 4.2: The outcome frequencies after 200 rounds of the IPD

The cooperative outcome (CC) occurred only 31.5%. The agents did not follow a particular strategy and chose their action in a random non-reward-maximizing manner. The CD outcome occurred 21.5%, the DC 26% and the DD 21%.

In effect, these opposite in sign signals update the value of the action that was not chosen by each network and can be justified as an additional feedback to the agents for their performance in the previous round. Overall during a learning round, each network receives global, constant and opposite in sign reinforcements that are applied in the timestep following spikes of both of its output neurons. One of the two signals is due to the payoff matrix of the game and its purpose is to "encourage" or "discourage" the action that elicited reward or penalty and the other signal is complementary and is purpose is to "encourage" or "discourage" the action that could have elicited reward or penalty if it had been chosen in the previous round of the game. Table 4.1 summarizes all the administered reinforcements.

	Output 1	Output 2	Output 3	Output 4
CC	+1.4	-1.15	+1.4	-1.15
CD	-1.3	+1.15	-1.15	+1.5
DC	-1.15	+1.5	-1.3	+1.15
DD	+1.15	-1.2	+1.15	-1.2

Table 4.1: Overview of the reinforcement signals

The table summarises the reinforcement signals (as applied on the equations) which the two networks receive during a learning round, according to all possible outcomes of a given round of the game. The reinforcement is administered for every spike of output neurons *1* to *4* (see figure 3.1 for output neuron numbering).

Figure 4.3 and 4.4 show the system's performance when additional reinforcement signals were incorporated into the learning algorithm. The simulation was identical to the previous one apart from the enhanced reinforcement administration scheme. The difference in performance is evident. The networks accumulated a total payoff of almost 1500 by cooperating 91% of the times. The results reveal that the agents learned to maximize long term reward through cooperative behaviour. According to the interpretation of the game the two subagents of the brain managed to engage in compromising behaviour and achieve self-control. It has to be noted that the CC outcome not only persisted during the final rounds of the simulations, but it also did not change after a point due to the system's dynamics that were evolved by that point in time in such a way to produce CC consistently. Unless specified the following simulations are carried out with extra reinforcement administration.

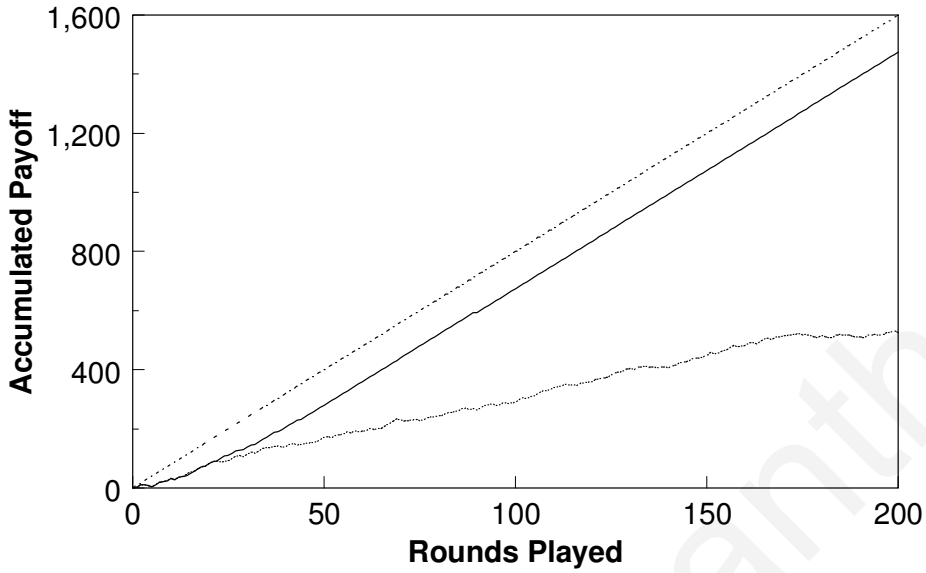

Figure 4.3: Simulating internal conflict through reinforcement of stochastic synaptic transmission

The system's performance during the IPD with (*solid line*) and without (*dotted line*) the extra reinforcement administration. The performance increased dramatically when extra global signals were given as a feedback to the agents. The agents managed to engage in mutual cooperation and therefore exhibited self-control behaviour. The theoretically best performance is shown for comparison (*dot-dashed line*).

As explained in Chapter 3, the eligibility trace is a dynamical variable used to integrate time related evens and is utilized in the current algorithm as a memory for each synapse's past actions with respect to releasing a neurotransmitter. The eligibility trace time constant regulates the decay of the variable and signifies for how long these events affect the variable regulated by learning (i.e probability of the neurotransmitter release in this case). In other words, a synapse with greater eligibility time constant has a stronger memory on its past actions than a synapse with smaller eligibility time constant, and employs this memory in order to decide whether to increase or decrease the

neurotransmitter release probability given the reinforcement it receives. A neuron equipped with such synapses can therefore be considered to have a memory on these actions and so a network with such neurons can be considered to have a memory on these actions as well.

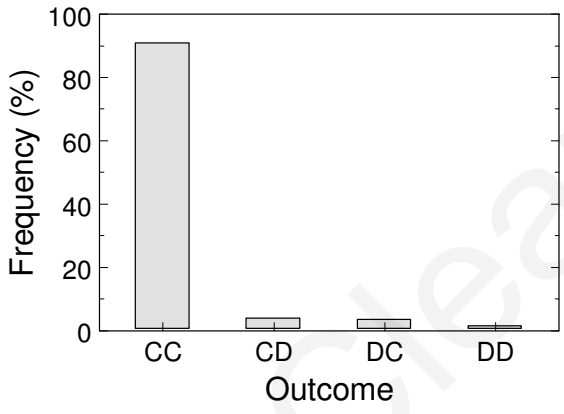

Figure 4.4: Game percentage outcomes with extra reinforcement. The outcomes after 200 rounds of the IPD. The cooperative outcome (CC) was successfully learned and occurred most of the times (91%). The other outcomes took place at the beginning of learning in small percentages. CD occurred 4%, DC 3.5% and DD 1.5%. The differences in the percentages respective to the ones obtained without extra reinforcement (Figure 4.3) are statistically significant using a one-tailed z-test at 95% confidence interval.

Therefore, the network employs this memory through its synapses in order to maximise the reward and so we can claim that the netwotks have reward-directed memory. The effect is the same as if the networks had a memory of the actions of cooperation and defection with respect to the reward accumulation and decided whether to cooperate or defect such that they maximised reward. For these reasons, a network comprising synapses with high eligibility trace time constants is used to implement a

subagent with stronger memory and a network with low eligibility trace time constants to implement a subagent with weaker memory.

The following simulations are carried out in order to investigate the effect of the networks' memory on attaining the cooperative behaviour and thus the effect of the agent's memory on achieving self-control behaviour.

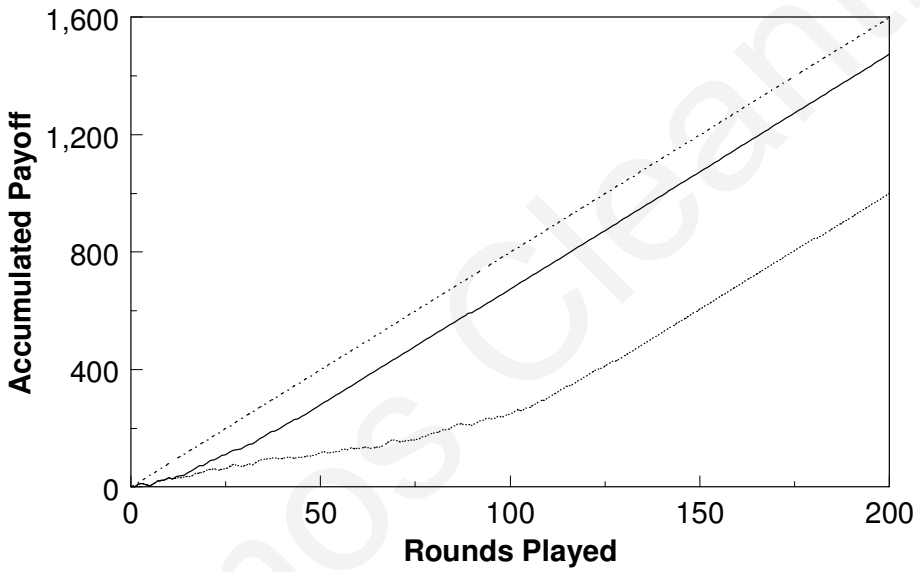

Figure 4.5: The eligibility time constant effect

The eligibility trace time constant effect (with extra reinforcement) when the spiking NNs learn with reinforcement of stochastic synaptic transmission. The system collected a much higher total reward when the eligibility trace time constant of both networks was equal to 20ms (*solid line*) compared to 2ms (*dotted line*). The theoretically best performance is shown for comparison (*dot-dashed line*).

Two simulations were performed with the synapses of the two networks having different eligibility trace time constants. The values for both networks were set to 20ms and 2ms for the two simulations respectively. Therefore, during the first simulation both networks have a "strong memory" whereas in the second they have a "weak memory".

The performance of the system for all simulations is shown in Figure 4.5. The difference in the system's performance is obvious and significant. When the system was configured with 20ms eligibility trace time constants, the accumulated payoff is much higher than the one with 2ms; this results from the difference in the cooperative outcome. With the eligibility trace time constants set at 20ms the two networks learned quickly to cooperate in order to maximise their long-term reward and achieved the CC outcome 182 out of the 200 times. On the contrary, when the system was configured with "weak memory", learning took effect much later during the game (after the 100th round) and thus the system exhibited much less cooperation (120 out of 200). However, the system with both configurations eventually managed to learn how to cooperate.

Results show that networks' memory influences the cooperative outcome of the game in the sense that it could delay it to a great extent. However, a weak memory does not destroy learning as the networks eventually learned to cooperate. On the other hand, the administration of extra reinforcement was vital for learning the desired behaviour, no matter the memory strength of the agents.

.

4.3 Reward-Modulated STDP with Eligibility Trace

The following simulations implement the IPD where the agents learn through reward modulated STDP with eligibility trace (Florian, 2007). The simulations aim to investigate the capability of the spiking NNs to cooperate in the IPD or equivalently, under the interpretation of the game, to investigate the capability of the simulated subagents to exhibit self-control. It is noted that unless specified the learning rate used in the simulations is 0.7×10^{-4}. The learning rate must be very small since the networks are reinforced for every spike of their output neurons. Therefore a small learning rate

combined with small reinforcement signals ensure that changes on the variables controlled by learning are made in a smooth and gradual way.

As shown in the previous simulations, the administration of additional, opposite in sign, global reinforcement signals proved to be vital for the successful training of the competing agents that attained the cooperative outcome. We therefore tested the importance of this additional reinforcement administration, for the performance of the system when trained with reward-modulated STDP with eligibility trace (Florian, 2007). Figure 4.6 shows that the implementation of the game was successful when the additional reinforcement signal was administered.

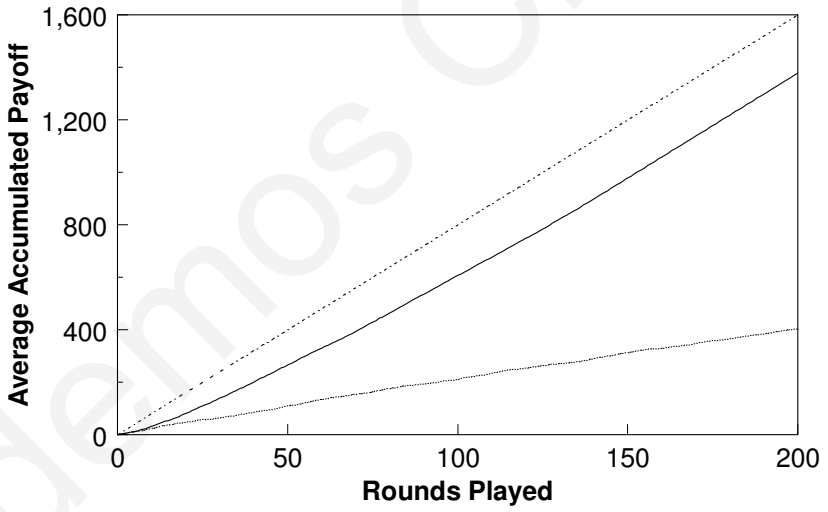

Figure 4.6: Simulating internal conflict through reward-modulated STDP with eligibility trace: the effect of the extra reinforcement administration

The system performed much better when extra global reinforcement signals were given as a feedback to the agents (*solid line*). In contrast, it accumulated a very small total payoff when no additional signals were given (*dotted line*). The theoretically best performance is shown for comparison (*dot-dashed line*).

The cooperative outcome was often attained after a relatively short training period, which enhanced the accumulation of reward by the system. This reveals that after a certain point the networks successfully learned to resist the temptation payoff provided by defection in order to maximise their long-term reward through cooperation, enabling thus reward maximisation by the system as well. The system's performance corresponds to the conflict being solved by self-control behaviour. However, the system performed badly when no extra reinforcement was given.

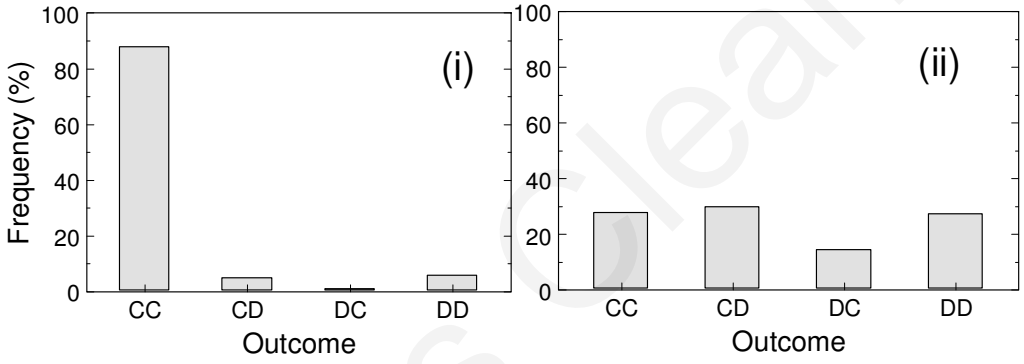

Figure 4.7: Game percentage outcomes with extra reinforcement (i) vs. no extra reinforcement (ii)

Eligibility trace time constant is set to 25 ms for both simulations. The cooperative outcome (CC) was satisfactory learned and occurred 88% of the times in the case where additional reinforcement was signaled to the agents. The other outcomes resulted in small percentages. (CD=5%, DC=1%, DD=6%). With no extra reinforcement mutual cooperation (CC) occurred only 28%, similar to the CD and DD case (30% and 27.5 respectively) whereas DC outcome was at 14.5%. The differences in the percentages are statistically significant using one-tailed z-tests at 95% confidence interval.

Figure 4.7 shows the outcomes obtained in the game during the two simulations. The agents cooperated 88% of the times when the extra reinforcement was introduced.

The performance deteriorated significantly when no additional reinforcement signals were administered to the networks since the cooperation level fall 60 percentage units (from 88% to 28%) and the defection level increased 21.5 percentage units (from 6% to 27.5%). The results with the current learning scheme are in line with our previous results with regards to the effectiveness of the additional reinforcement in the attainment of a cooperative behaviour. The administration of extra reinforcement is thus vital for a high accumulated payoff by the spiking NN agents; therefore all the subsequent simulations are carried out with extra reinforcement administration.

In reward-modulated STDP with eligibility trace, the latter serves as a decaying memory of the relation between recent pre- and postsynaptic spike pairs. Its time constant signifies the length of time that a given event (in this case a spike pair) affect the variable regulated by learning (in this case the synaptic strength). By applying the same reasoning as in section 4.2, a network comprising synapses with high eligibility trace time constants is used to implement a subagent with stronger memory and a network with low eligibility trace time constants to implement a subagent with weaker memory.

The following simulations are carried out in order to investigate the effect of the networks' memory on attaining the cooperative behaviour and thus the effect of the agent's memory on achieving self-control behaviour. Three simulations were performed with the neurons of the two networks having different eligibility trace time constants. The values for both networks were set to 25ms and 2ms respectively for the two simulations, whereas during the third one, one network was configured with 25ms and the other with 2ms. Therefore, during the first simulation the agents had a strong memory, in the second they had a weak memory and in the third one agent had strong and the other had weak

memory. The performance of the system for all simulations is shown in Figure 4.8 and the respective outcomes are shown in Figure 4.9.

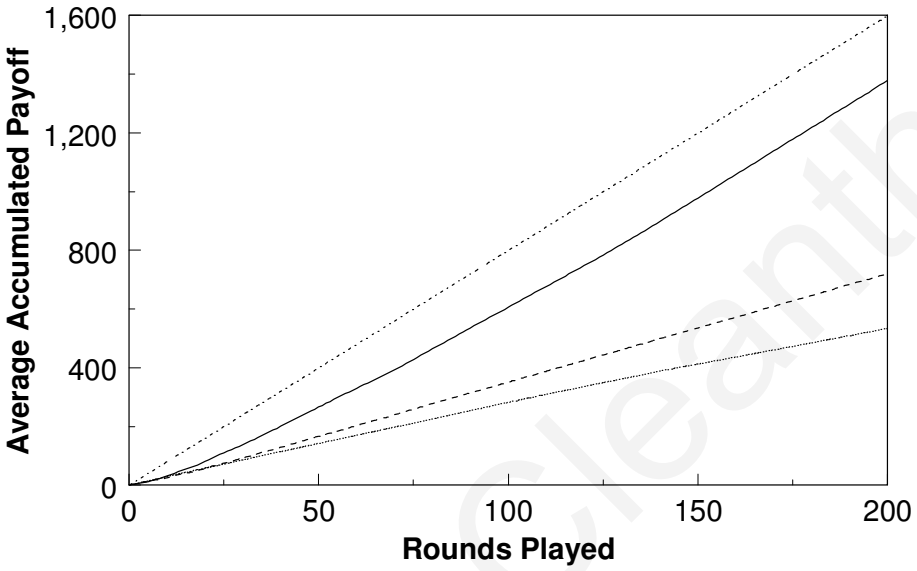

Figure 4.8: The eligibility trace time constant effect with reward-modulated STDP

The system collected a much higher total reward when the eligibility trace time constant of both networks was equal to 25ms (*solid line*) compared to 2ms (*dotted line*). The system performed in between when one network was configured with 25ms and the other with 2ms (*dashed line*). The theoretically best performance is shown for comparison (*dot-dashed line*).

The difference in the system's performance is evident. When the system was configured with 25ms eligibility trace time constants, the accumulated payoff was much higher than in the case where the system was configured with 2ms eligibility trace time constants. During the former simulation, the agents engaged in a behaviour of mutual cooperation whereas in the latter they primarily defected. With the eligibility trace time constants set at 25ms the two networks learned quickly to cooperate in order to maximise their long-term reward and achieved a total payoff of 1379 with the CC outcome chosen

88% of the times. On the contrary, when the system was configured with weak memory (2ms eligibility trace time constant for both agents), the system exhibited much less average cooperation (50%) and a total payoff of 534.

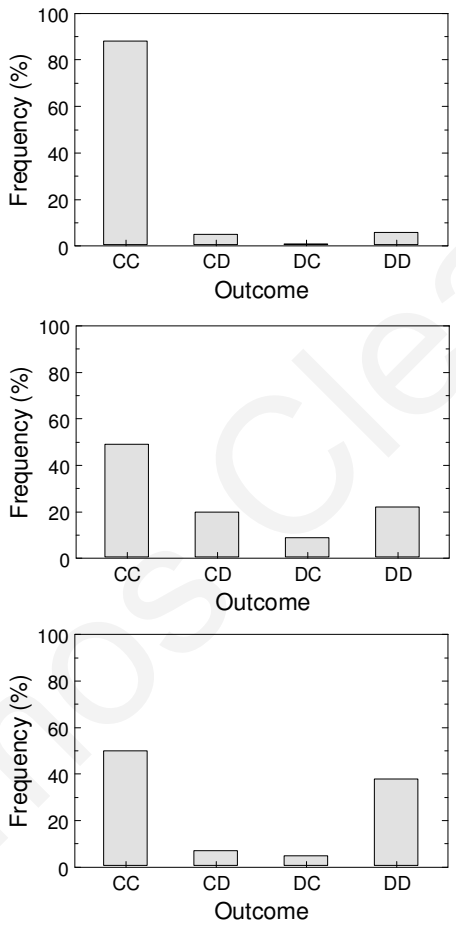

Figure 4.9: Game percentage outcomes

Average outcomes after 200 of the IPD when both networks have a strong memory (Top), weak memory (Bottom) and when one has strong and the other weak (Middle). Eligibility trace time constant is set to 25 ms for a strong memory and 2ms for a weak memory. The cooperative outcome (CC) was satisfactory learned and occurred 88% of the times in the case of strong memory agents whereas cooperation level diminished significantly (also statistically significantly using a one-tailed z-test at 95% confidence interval) when one of the agents or both had a weak memory (49% and 50% respectively).

The system performed slightly better in the final simulation where one network had a strong memory and the other had a weak one. It accumulated a total payoff of 720 compared to the 534 of the "memoryless" networks. However, the cooperation remained at the same low level (49%).

The difference in the accumulated payoff occurs due to the difference in the DD and CD outcome rather than in the CC outcome. In the case of strong memory vs. weak memory, increased CD outcome reveals that the strong memory agent was trying to engage in a cooperative behaviour by playing C but the weak memory could not 'realise' and thus chose to aim for the temptation payoff by playing D. Later on the strong memory agent adjusted and changed its strategy by playing D as a best response to the other agent. In the case where both agents had a weak memory they both aimed for the temptation payoff and thus engaged in a behaviour of stronger mutual defection (DD outcome was 38%). Only the system with the strong memory configuration managed to exhibit high cooperation levels.

4.4 Further Investigation of the Performance of Reward-Modulated STDP with Eligibility Trace: Does High Firing Irregularity Enhance Learning?

In Section 3.5 of Chapter 3 we made the hypothesis that high firing irregularity at high rates enhances learning. We showed that when LIF neurons fire highly irregularly at high rates then, the performance of a spiking neural network (trained with reward-modulated STDP (Florian, 2007)) is significantly better. The high firing irregularity at high rates was achieved by the use of the partial somatic reset mechanism on every LIF neuron of the network. We believe that this is due to the high irregular firing achieved by the LIF neurons that enabled the algorithm to perform more accurate correlations between pre-

synaptic and postsynaptic spike timings and reinforcement signals (for more details please see Section 3.5).

In this section we carry on by testing our hypothesis in a much more complex MARL task such as the IPD. As in Section 3.5 we made sure that prior learning, the output neurons in both systems, with and without partial somatic reset, fire at the same experimentally observed high frequency of 100 Hz, by having different input frequencies for the two systems. This was done to ensure that any difference in the performance is due to the high irregularity enabled by partial reset and not to the increased output firing that would also be enabled by partial reset. It is noted that the output firing rate was influenced by learning in the duration of the experiments, but not to a great extent and in the same manner for the two systems throughout the simulations.

The results of both simulations in the IPD multiagent RL task are shown in Figure 4.10. With both configurations the system learns to cooperate, but when each of the competing networks of the system comprises of LIF neurons equipped with the partial somatic reset mechanism, the accumulated payoff is much higher than when there is total reset after each firing spike; this results from the difference in the cooperative outcome. With the partial reset the two networks learned quickly to reach very strong cooperation in order to maximise their long-term reward and achieved the CC outcome 61% of the time on average. On the contrary, with total reset, learning is not as strong, which is evident by the fact that the system exhibited much less cooperation (39% of the time on average).

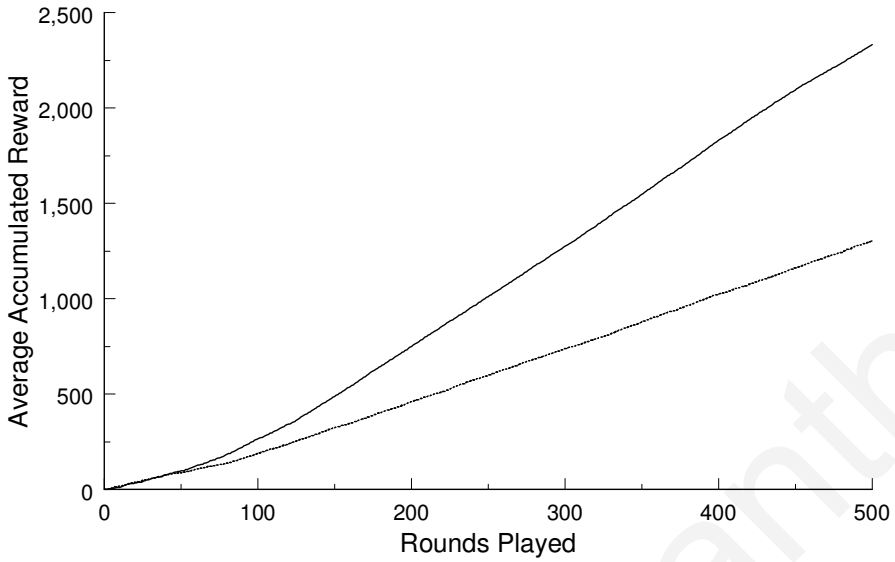

Figure 4.10: Effect of increased firing irregularity on the performance of the system when implementing the IPD

Average accumulated reward with the LIF neurons of both networks having either partial somatic reset at 91% of threshold (solid line) or total reset (dotted line). For both networks the eligibility trace time constant τ_z is set to 25ms and the learning rate to 0.0007.

4.5 Discussion

4.5.1 Internal Conflict and Self-Control Behaviour

Overall, results obtained by both employed algorithms show that self-control behaviour can be learned. The simulated internal agents achieved through learning to postpone immediate gratification in favour of a superior outcome in the long run. Our results are in line with the analytical work of Livnat and Pippenger (2006), who showed that an optimal brain can be composed of internal competing agents. Optimality in the case of internal conflict is the behaviour of self-control that can be attained if the internal agents follow a compromising strategy in their interaction. Therefore our results show that self-control

behaviour could be a learned maximising strategy employed by a reward maximising brain in the presence of competing internal agents.

In addition, results show that for individuals that experience a particular internal conflict in a recurrent manner, self-control behaviour can be learned through correct associations between actions and outcomes. More specifically, an individual that experiences the same internal conflict for a number of times, is more likely to learn how to practise self-control on the specific matter as the recurrent experience will enable the individual to associate the actions aiming at immediate gratification with the suboptimal obtained outcome (DD). In other words the individual will learn that the best outcome cannot be obtained by insisting on acquiring the best immediate available option of either competing 'self' as by insisting can cause future unavailability of that option (CD or DC can cause an extended DD outcome due to punishment by the other 'self') or can result in an unintended immediate outcome if both of the subagents insist on getting their way (DD).

In addition, recurrent experience enables the endurance of self-control behaviour once established, since it allows the system to acquire an appreciation of the accumulated reward obtained by executing self-control behaviour consistently. In our computational model the appreciation of the accumulated reward is reflected in the consistent activation of the specific output units such that to enable persistent mutual cooperation. Therefore, in addition to experimental findings that showed a greater fronto-parietal activity in subjects who chose long-term rewards in intertemporal choice tasks (e.g. McClure et al., 2004), our results suggest that people who consistently and efficiently practise self-control in their everyday lives should have a greater respective neural activity induced by

the appreciation towards the overall accumulated reward resulted from consistent practice of self-control behaviour.

Simulations revealed that the reward-correlated memory of the competing agents, facilitated by the variable of eligibility trace and its time constant, is also important in the process of learning self-control behaviour since the memory is employed by the agents in choosing their actions such that to maximise their individual reward. Agents with stronger memory were able to learn fast that mutual cooperation elicits the highest long-term payoff and adjusted their strategies such that to collect it. On the other hand, agents with very low memory were not always able to learn to cooperate and in the cases they did, it took longer. In addition, they demonstrated high percentages of mutual defection revealing an 'eagerness' to collect the immediate payoff and an inability to learn that defection does not pay off in the long term; the agents behaved in a myopic manner. Finally, in the simulation where one agent had strong memory and the other had weak, increased frequency of CD outcome shows that the one with the strong memory was exploring cooperation while the other primarily defected. In this situation, the one with the shorter memory accumulated a higher total individual payoff due to collection of the Temptation payoff. The short-memory agent exploited the strong-memory agent until the latter also changed its strategy to defection.

If we assume that some internal subagents are more myopic than others (distinguished by the time they usually wait in order to collect the rewards), then results suggest that myopic subagents (e.g. satisfy-hunger or taste agent) are more likely to exploit subagents that are less myopic and are willing to compromise (e.g. physical fitness agent). This might provide a psychological plausible picture of why we sometimes excess on sweets while dieting does not begin on Mondays.

Overall, results with respect to the effect of eligibility trace time constant show that strong reward-directed memory is important for the attainment of self-control behaviour. As in our computational model, this kind of memory might not concern the actual actions performed by the individual, but it might be implicitly present in the organism for optimization purposes. However its reward-driven nature enables maximisation by the individual through the performed actions as if the individual had an explicit memory of these actions.

4.5.2 High Firing Irregularity and Learning

In general, the findings from our experiments in the XOR and the IPD tasks, suggest that the increased firing irregularity at high rates, which results from the introduction of the partial somatic reset mechanism at every LIF neuron of these networks, enhances the learning capability of both systems. This is due to the increased suppression of the output firing rate for input pattern {1, 1} in relation to the output firing rates for input patterns {0, 1} or {1, 0} in the XOR problem and the resulting accumulation of higher cooperative reward in the IPD task. More specifically, this high firing irregularity at high rates enhances reward-modulated STDP with eligibility trace. We believe that this is due to more accurate correlations between pre-synaptic and postsynaptic spike timings and reinforcement signals. If firing is regular, then it is possible for two identical spike pairs to be associated with opposite in sign reinforcement signals, confusing thus the direction of the plasticity for a given synapse. High firing irregularity prevents this unnecessary competition by weakening this possibility and thus preventing a possible corruption of the learning algorithm. We have also observed that the increased levels of temporal

irregularity only have 'positive' effects, because they either increase the speed in a successful learning episode, or reverse a failed learning episode into a successful one.

It has to be noted that other variant implementations of RL on spiking neural networks by modulating STDP with a reward signal (apart from Florian, 2007), like for example Izhikevich (2007), Faries and Fairhall (2007) and Legenstein, Pecevski and Maass (2008), could equally well be used for obtaining the results presented in this research thesis. In general, the use of LIF neurons with the partial somatic reset mechanism is very important, as apart from its precise modelling of the high firing irregularity of cortical neurons at high firing rates (Bugmann, Christodoulou and Taylor, 1997; Christodoulou and Bugmann, 2001), it also enhances learning.

The results regarding learning and high firing irregularity are indeed important and have been accepted for publication in a relevant scientific journal (Christodoulou and Cleanthous, 2010). However, no direct comparisons can be made between results presented in sections 4.3 and 4.4 as simulations in Section 4.4 aimed at a different output firing that was enabled with different input frequencies for the two systems. For subsequent simulations we will not employ partial somatic reset on the LIF neurons because such biological realism is superfluous in the scope of this research thesis as the output units in the system correspond in an abstract way to aggregate activation by neural systems involved in internal conflict and not to a particular neuron as such.

Chapter 5

Exploring the Structure of Internal Conflict

5.1 Overview

In the previous chapter we simulated internal conflict as a competitive interaction between rational subagents where the conflict structure was represented by the payoff matrix of the IPD. All the simulations used a particular structure of the payoff matrix where the payoffs of the competing agents were symmetric and constant throughout the simulations. In the current chapter we explore how the structure of internal value conflicts, as represented by the payoff structure of the game, influences the attainment of self-control behaviour. We attempt this by constructing a variety of payoff matrices that correspond to different internal conflict contexts and apply them to our computational model in order to investigate how they affect the cooperative outcome. Experiments employ constant payoff structures that do not change during the duration of the game as well as varying payoff structures in order to simulate time related changes in the value systems of internal agents. All simulations presented in this chapter were performed by employing reward modulated STDP with eligibility trace and the results.

5.2 Simulating Internal Conflict Scenarios with Constant Payoff Structures

The previous results involved a payoff matrix (Table 3.1) which represented an internal conflict of low to moderate intensity. This is so because the Temptation payoff is just slightly higher than the payoff for mutual cooperation and also because the Sucker's payoff is slightly lower than the payoff for mutual defection. In other words, the agents were less tempted to defect and less afraid to cooperate. As we showed, internal agents competing in this context were able to learn to exhibit self-control by compromising and therefore accumulate superior long term reward. How would the agents respond if the conflict was more intense? In order to answer this question two more sets of experiments were performed with the two new payoff structures shown below (together with old one for comparison). For simplicity, payoff matrices in Figure 5.1 contain the reinforcement signals as applied directly on the equations of the learning algorithm. Payoff matrix (i) in Figure 5.1 is the same as the one used in all simulations of Chapter 4 and is presented here for comparison to the new ones.

Both the Temptation and the Sucker's payoff of Agent *II* (column player) were modified accordingly in order to reflect a situation where an individual experiences a greater internal conflict. The Temptation payoff reflects how much the immediate gratification outcome yields whereas the Sucker's payoff signifies the cost of obtaining the least immediate gratification. The Temptation payoff was increased from 1.5 to 7 for the one experiment and then to 14 for the other whereas the Sucker payoff was decreased from -1.3 to -5 and then to -12. As a result, the agents' payoffs are more divergent for the CD and DC outcomes increasing thus the conflict between them.

(i)	Agent II		
		Cooperate	Defect
Agent I	Cooperate	1.4, 1.4	-1.3, 1.5
	Defect	1.5, -1.3	-1.2, -1.2

(ii)	Agent II		
		Cooperate	Defect
Agent I	Cooperate	1.4, 1.4	-1.3, 7
	Defect	1.5, -5	-1.2, -1.2

(iii)	Agent II		
		Cooperate	Defect
Agent I	Cooperate	1.4, 1.4	-1.3, 14
	Defect	1.5, -12	-1.2, -1.2

Figure 5.1: Payoff matrices representing different intensities of internal conflict

The top matrix (i) is the original matrix used in the simulations of Chapter 4 and corresponds to an internal conflict of moderate intensity whereas the second (ii) corresponds to a strong internal conflict and the third (iii) to an extreme conflict similar to that experienced in cases of addiction. Differentiation in conflict intensity is implemented by modifying the payoffs in the CD and DC outcome for Agent II. Payoffs are shown as applied on the equations of the learning algorithm.

Moreover, the modifications in both directions promote the choice of defection for the column player as the individual is more tempted to go for the greater immediate payoff by defecting as well as to avoid the Sucker payoff by also defecting. Payoff matrix (ii) and (iii) of Figure 5.1 model a strong and an extreme internal conflict respectively.

The payoffs for the column player in the second matrix are so high compared to the original matrix that we could think of it as representing an extreme case of value system similar to that of addiction. Another point to note is that the game still obeys the rules of the Iterated Prisoner's Dilemma despite the fact that the payoffs for the two players are not symmetric anymore.

The results for the original payoff matrix as well as for the two new ones are shown in Figure 5.2. In general, the system behaved as anticipated. The column player indeed defected more in the strong internal conflict scenario and defected even further in the extreme conflict scenario as reflected in the increased percentage of the CD outcome. As a response to this, the row player also learned to defect in order to secure the payoff gained from mutual defection, which is better than the Sucker's payoff, increasing thus the DD outcome. Notice that the row player learned to efficiently adopt its strategy in the face of a new strategy by the column player showing thus flexibility of the system. As a result of the above the cooperative outcome decreased significantly. More specifically the CC outcome was 86% during the simulation that modeled an internal conflict of low intensity and decreased to 50% during the strong conflict while it decreased even further to 30% during the extreme conflict whereas the CD outcome increased from 5% to 25% and then to 30% respectively. The DD outcome also increased from 4% to 25% and then to 40%. It is important to note that the synaptic changes during the extreme conflict scenario were very drastic and persisting due to the very high magnitudes of the new payoffs, and as a result they hindered learning to a great extent even from the initial stages of the simulation. These changes might relate to the extreme and persisting neuroadaptations caused by addictive substances but more on that subject will follow in Section 5.3.

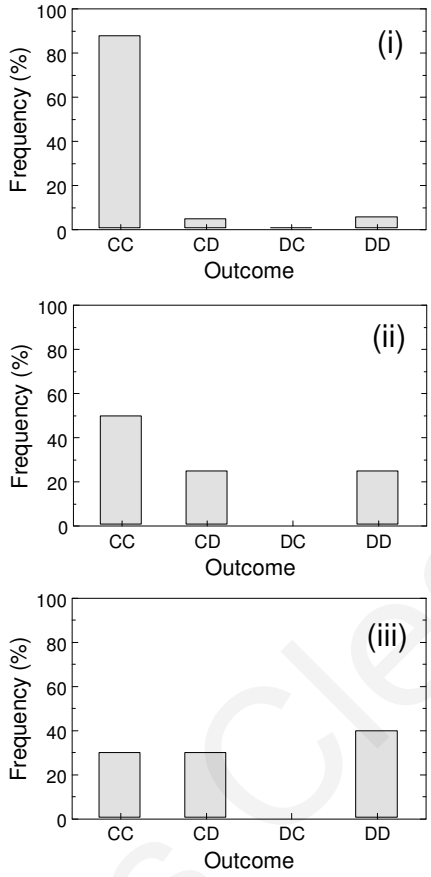

Figure 5.2: IPD outcomes for different intensities of internal conflict

Average outcomes after 200 rounds of the IPD simulating a (i) moderate, (ii) strong and (iii) extreme internal conflict. Outcomes for (i) are CC= 88%, CD= 5%, DC= 1%, DD= 6%, for (ii) are CC= 50%, CD= 25%, DC= 0%, DD= 25% and for (iii) are CC= 30%, CD= 30%, DC= 0%, DD= 40%. It is noted that apart from the payoff matrix, the system for the three simulations is configured in an identical manner. Learning rate is set to 0.00007 and the eligibility trace time constant is set to 25ms. The decrease in the CC outcome from case (i) to cases (ii) and (iii) is statistically significant using a one-tailed z-test at 95% confidence interval whereas the increase in the DD percentage from case (i) to (ii) and from (ii) to (iii) is also statistically significant using a one-tailed z-test at 95% confidence interval.

The previous two simulations represented conflict between value systems of different importance to the individual as reflected in the magnitudes of the payoffs that

each outcome yields. Overall, Agent *I* expects small gains and losses from the different outcomes compared to Agent *II* who expects greater gains and losses. Therefore, the value system served by Agent *II* is more important to the individual. For the purposes of the following simulations we modified the Temptation as well as the Sucker's payoff for Agent *I* (row player) in the exact same manner as we did for Agent *II* in the previous two simulations. The new payoff matrices are shown in Figure 5.3. As a result, the following simulations implement two agents with equally important value systems (symmetric payoffs) and an increased motivation to defect.

(i)	Agent II		
Agent I		Cooperate	Defect
	Cooperate	1.4, 1.4	-5, 7
	Defect	7, -5	-1.2, -1.2

(ii)	Agent II		
Agent I		Cooperate	Defect
	Cooperate	1.4, 1.4	-12, 14
	Defect	14, -12	-1.2, -1.2

Figure 5.3: Payoff matrices when simulating intense internal conflict between two agents who are strongly motivated to defect.

(i) Both agents have Temptation and Sucker's payoff equal to 7 and -5 respectively. (ii) Both agents have Temptation and Sucker's payoff equal to 14 and -12 respectively. Payoffs for mutual cooperation and defection are 1.4 and -1.2 respectively for both agents on both payoff matrices.

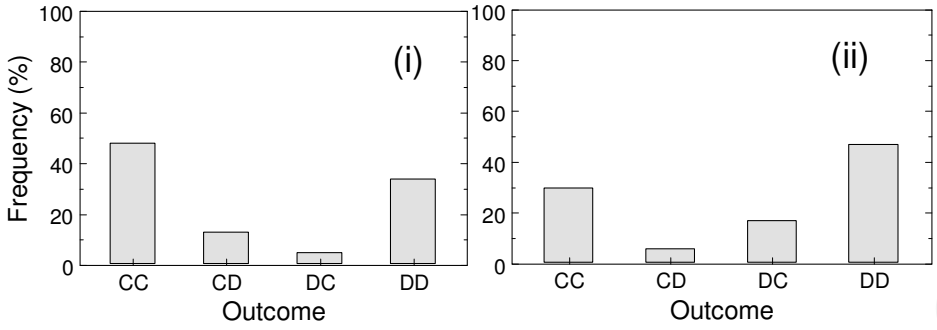

Figure 5.4: IPD outcomes when simulating intense internal conflict between two agents who are strongly motivated to defect.

Average outcomes after 200 rounds of the IPD when (i) both agents have Temptation and Sucker's payoff equal to 7 and -5 respectively and (ii) the agents have Temptation and Sucker's payoff equal to 14 and -12 respectively. Payoffs for mutual cooperation and defection are 1.4 and -1.2 respectively for both agents on both simulations. When both agents have equal strong motivation to defect then the bigger the motivation (as reflected in higher Temptation and lower Sucker's payoff), the lower the cooperative outcome and the higher the mutual defection. The decrease in the CC outcome as well as the increase in the DD outcome from case (i) to (ii) is statistically significant using a one-tailed z-test at 95% confidence interval.

Results presented in Figure 5.4 show the outcomes of the game after 200 rounds of the IPD (Subfigure (i) corresponds to the simulation with the Payoff matrix (i) of Figure 5.3 and Subfigure (ii) corresponds to the simulation with the Payoff matrix (ii) of Figure 5.3). Overall, results show (Figure 5.4) that when both agents have equal strong motivation to defect then the bigger the motivation (as reflected in higher Temptation and lower Sucker's payoff), the lower the cooperative outcome and the higher the mutual defection. Results are more interesting when compared to the case where only one of the agents was motivated to defect. Results in Figure 5.4 show that both agents could not

escape temptation and they both insisted in getting their way, resulting thus in a behaviour of increased mutual defection compared to the respective cases (Figure 5.2) where only one of the agents had strong motivation to defect. The cooperative outcome was not influenced to a great extent; the increase in DD is due to a decrease in CD rather than a decrease in CC. DC outcomes also increased compared to Figure 5.2 as Agent *I* has also strong motivation to defect. However they remain in low levels due to presence of another "defector". Note also that given that the CC outcome stayed at the same levels and DD is more costly to the system than CD or DC, then we can infer that the system is worse off when both agents are strongly motivated to satisfy maximum immediate gratification rather than when only one of the subagents is motivated even though strong mutual cooperation is not attained in any of the two cases. Therefore, one could say that it is more costly for the individual to experience an internal conflict between two value systems that are equally important with large respective values attached to immediate gratification rather than when only one of the value systems is as such, even though self-control behaviour is not attained in any of the two cases.

The results reveal that the structure of the payoff matrix is highly important for the outcome of the game. In addition, it is shown that the payoff matrix of the IPD is very powerful in abstractly representing complex settings with respect to an individual's internal conflict and in addition the system can efficiently exploit this powerful representation in order to simulate an individual's respond to the intensity of internal conflict. Now given that we achieved to simulate a situation where an individual faces a strong internal conflict, results showed that learning solely is not sufficient to overcome the conflict in a self-controlled manner, especially in cases where conflicting desires or value dimensions are equally important and immediate gratification yields much more

than the single self-control outcome (CC) (not self-control behaviour which is the consistent and persistent choice of CC). Instead, the simulated individual is likely to give-in to temptation or even more likely, according to Kavka's (1991) interpretation, to preserve the *Status Quo* (DD outcome) and thus satisfy neither of the two internal subagents.

Considering the previous results, it would be interesting to see how cooperative behaviour would change with respect to the other outcomes, by 'boosting' appreciation towards the cooperative outcome. This was done during the following simulation by using the payoff matrix presented in the table below (Table 5.1). It is the same as the one used in one presented in Figure 5.3 (i), but with increased payoffs for the outcome of mutual cooperation.

		Agent II	
		Cooperate	Defect
Agent I	Cooperate	6, 6	-5, 7
	Defect	7, -5	-1.2, -1.2

Table 5.1: Payoff matrix with equally important value systems and increased mutual cooperation payoff

The table summarises the context of internal conflict. It is an intense conflict due to the great difference between the individual payoffs of the agents in the CD and DC case but with also increased payoffs for mutual cooperation so that appreciation for mutual cooperation can be build up easily.

Results presented in Figure 5.5 (ii) show that the system managed to cooperate more compared to the case where the payoffs for mutual cooperation were low (shown in Figure 5.5 (i) for comparison, corresponding to the Payoff matrix (i) of Figure 5.3). In addition, the difference results mainly from a decrease in the DD outcome rather than from a decrease in CD or DC outcome which is more beneficial for the system, as a CD or DC outcome provides a positive overall reinforcement whereas DD a negative. Once more the system behaves in a way that makes sense, and therefore reinforces our idea that building up appreciation for the cooperative outcome can be beneficial in cases of intense conflict.

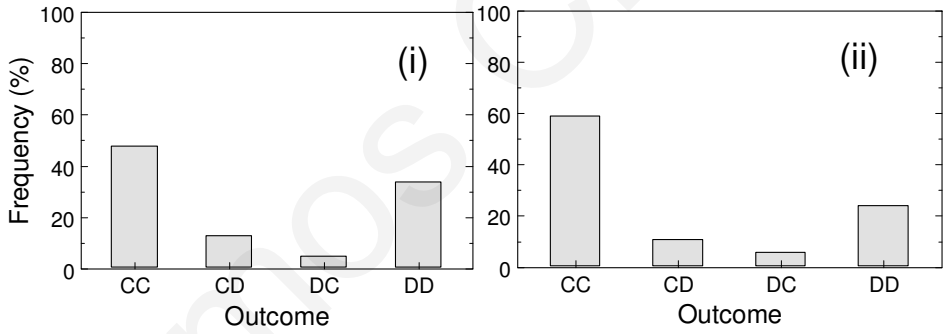

Figure 5.5: IPD outcomes when simulating intense internal conflict with equally important value systems but increased mutual cooperation payoff

Average outcomes after 200 rounds of the IPD when (i) mutual cooperation payoffs were low and (ii) mutual cooperation payoffs were high. CC percentage increased compared to (i) from 48% to 59% while the DD percentage decreased from 34% to 24%. The respective increase and decrease are statistically significant using a one-tailed z-test at 95% confidence interval.

Although this kind of internal value structure is preferable and beneficial to the individual, compared to the one that reflects intense conflict but with low immediate self-control payoff, it is also more rare in real life situations of internal conflict. Usually

people can not appreciate self-control behaviour from single self-control actions in the presence of big immediate temptations, because of the small immediate payoff provided by self-control. These small self-control payoffs combined with high temptation payoffs limit their ability of practicing self-control long enough such that an appreciation is built up. In such a case, is there a way of inducing an appreciation for self-control behaviour early on, without having to exercise self-control too many times? As we saw, that would mean that the payoffs delivered from exercising self-control should for some reason increase dramatically. But how can the payoffs increase if the outcome delivering these payoffs remains the same? One plausible hypothesis is that payoffs could increase not in their objective value as such, but in the way they are perceived by the individual. Our results are supported by findings (Metcalfe and Mischel, 1999) that showed that children can be taught to suppress impatience by manipulation of thought. Moreover, a recent study on students' delay of gratification (Bembenutty, 2009) also showed that the delay is enhanced with the use of self-regulated learning strategies like reminding themselves of their overall values and goals. In addition, such perceptual changes might be possible to be induced by drug administered psychiatric treatment but this is only a hypothesis. The importance of this result lies in the identification of an internal value structure that can improve self-control behaviour. The specification of how to attain such a structure is out of the scope this research.

5.3 Introducing Time in the Modelling of Internal Conflict: Simulations with Varying Payoff Structures

Although in the previous section we identified internal value structures which either support or hinder self control behaviour, one of the big questions persists; how can the simulated individual achieve self-control behaviour when experiencing a strong internal conflict, or at least improve the low self-control outcome of the previous simulations? Before we could move on we should first have a closer look at the current experimental procedure. So far, all simulations involve stationary implementations of the IPD in the sense that the payoff structures are fixed within the duration of a game. The two subagents compete for a given number of rounds in the IPD with a constant payoff matrix assuming thus fixed value systems. It is certain that such a constant representation of value systems for the two subagents is unrealistic in the sense that in real life the value systems shift all the time either because of internal or external causes. Consider our example where the student needs to decide whether to go to the pub or stay home and study. The conflict aroused by such a situation and its outcome are modelled by a single round of the IPD. In the course of its academic career, the student will face such a dilemma for an unspecified number of times, as modelled by the total rounds of the IPD. However it is not necessary that the student will have the same value systems every time s/he faces the dilemma. It is very likely that the student's value systems would have changed, during the time interval between two such situations. After all, time changes preferences. In addition, the value systems might alter on the spot as the decision is carried out. Imagine that as the student tries to decide what to do, s/he is informed that most of her/his friends are not going to the pub or s/he remembers a 'D' marked exercise on the course s/he has to attend the next morning. Both events will shift the student's

value systems. The first shift is due to an external cause initiated by the environment whereas the second one is due to an internal cause initiated by memory and emotions; however both causes enable the shift of the value systems.

The variable of time which always exists in real life scenarios and make the situations dynamic, although it is implicitly present in our computational system in the form of changes made to the variables affected by learning, or simply represented through time constants, it is absent in the representation of conflict itself. Therefore in order to integrate time in our computational model of internal conflict, the payoff matrix of the game should dynamically change simultaneously with the ongoing competition of the agents in the IPD. The value of such simulations would be to identify how alterations in the payoff structures, within the duration of a single game, affect the behaviour of the agents. Therefore we will be able to investigate how dynamic changes in the subagents' value systems might enable the transition from a non self-controlled behaviour to a self-controlled one.

Our first attempt to actively incorporate the effect of time was by applying the law of decreasing marginal utility (Gossen, 1854) on the payoff matrix of the game. This law comes from utility theory in economics and states that any additional unit of consumption of a good or a service yields less additional utility, a measure for satisfaction, than the previous unit. For example, imagine you have a bag of chocolates; the additional pleasure that you get from eating another chocolate is likely to be less than the pleasure received from eating the previous one. In an internal conflict scenario, a smoker who tries to decrease smoking for instance, enjoys the first cigarette of the day much more than the second and subsequent cigarettes as well as the prevention of smoking the first cigarette is much more painful than a subsequent one. In any case, in order to apply the law of

decreasing marginal utility in our model it would require that the Temptation payoff once consumed should induce a decrease to the Temptation payoff itself in the rounds that follow, and also an increase to the Sucker's payoff since the loss of not consuming the good would not be as severe now.

With regards to the simulation, the law of decreasing marginal utility was applied by gradually transforming the payoff matrix of the game from a very intense-conflict structure (Payoff matrix (iii) in Figure 5.1) to a low-conflict structure (Payoff matrix (i) in Figure 5.1) while the agents competed in the IPD. The game started with the intense-conflict matrix in the first round which was then transformed at the subsequent rounds, moving closer each time to the low-conflict matrix until it reached it at the final round of the game. As a result, the payoff matrix of the game changed, changing thus the agents' interaction setting, progressively from intense-conflict to low-conflict. The law of decreasing marginal utility is applied to all different goods and services, but at the same time the exact function that models this decrease is good/service-specific, and it differs from individual to individual. For these reasons a simple general linear transformation is chosen which captures the essential feature of the law, although many other functions could have been used. The purpose of the simulation was to establish whether the agents' behaviour would transform from a non self-controlled behaviour to a self-controlled one by the parallel ongoing transformation of the payoff matrix, given that the end matrix induced self-control behaviour when used in the static case.

Results obtained from this simulation were no different than the results obtained where the intense-conflict matrix was constant throughout the game (subfigure (iii) in Figure 5.2) and so there is no point in presenting them again. The linear transformation of the agents' preferences that took place along with the ongoing competition in the IPD was

not sufficient to alter their behaviour. The change in the agents' preferences or likenesses with respect to the possible outcomes, as reflected in the alteration of the payoff matrix of the game, did not induce a change in the agents wants as reflected in their respective decisions about the outcomes. We repeated the simulation where the starting matrix represented also an intense conflict but of a lesser magnitude (Payoff matrix (ii) in Figure 5.1), but again we observed no differentiation in the agents' performance.

In order to understand why the suboptimal behaviour induced by the intense conflict structure could not be reverted by a less intense payoff matrix, we should have a closer look at the training of the system. At the beginning of the simulation, the extreme value of the Temptation payoff and Sucker's payoff for Agent II induced a bias towards defection by the respective agent. The response of Agent I to the Agent's II defecting strategy would be either to Cooperate, something that would reinforce even more the action of defection by Agent II (since a CD outcome would still provide the best available reward for Agent II, despite the fact that the value of the Temptation payoff decreases as the game progresses), or adopt the best response to Agent's II strategy and Defect as well. In the latter case, where both agents Defect, one would expect that at some point in the game the agents would revert their strategies to mutual cooperation, as mutual defection would provide a constant penalty to both agents. However, Agent I is not biased towards the action of defection to the same extent as Agent II because of the much lower value of the Temptation payoff Agent I receives. Therefore, the synaptic changes required for Agent I to revert from defection to cooperation are much smaller compared to these required for Agent II. As a result, Agent I reverts to cooperation sooner than Agent II and the action of defection is again positively reinforced for the agents, especially for Agent II who receives once again the Temptation payoff (due to the CD outcome). Overall, this

cycle of behaviour prevents the agents from establishing a mutual cooperative behaviour despite the fact that the payoff matrix is transformed from intense-conflict to low-conflict.

One would expect that a change in likeness would affect a change in wants as we people usually act as rational utility agents. However this is not the case when it comes to intense cases of internal conflict as addiction. A person who is addicted to a given substance continues to consume that substance even in the absence of any pleasure initially provided by that substance (Fischman, 1989; Fischman and Foltin, 1992; Lamb et al., 1991). The dissociation between 'wanting' and 'liking' is central in modern addiction theory (Berger et al., 1996; Berridge and Robinson, 1998; Brauer and Dewit, 1996, 1997; Ohuoha et al., 1997; Robinson and Berridge, 1993) and provides justification for the 'irrational' choices made by addicts (Robinson and Berridge, 2000). According to incentive-sensitization theory (Robinson and Berridge, 2000) drugs not only produce long lasting changes to the neural systems "normally involved in the process of incentive motivation and reward", but also these changes make the reward systems sensitized (hypersensitive) to drugs and drug associated stimuli. According to this theory, the sensitised systems do not mediate the hedonic aspect of drugs (drug 'liking'), but instead they mediate a subcomponent of reward they refer to as incentive salience (drug 'wanting'). The initial euphoric effects of drugs induce neuroadaptations to the systems responsible for seeking that reward. Even in the absence of pleasure on a later stage, the adaptations made to seek the reward remain and guide the individual's actions.

Similarly in our computational model, the initial high reinforcement signals of the intense-conflict matrix induced great changes to the system's effective variables that drove the system in non self-control behaviour which persisted even when the payoff matrix changed. This is because the variables were affected to a greater extent by learning

at the initial stages of the game since the values of the Temptation and Sucker's payoff were much greater in magnitude. Thus, in analogy to the real neurobiological system we could say that our system was initially sensitised by the great Temptation and Sucker's Payoff towards the action of defection which persisted for the whole duration of the game irrespective of the payoff matrix used at the different stages of the game.

The next simulation aims at investigating the self-control strength model (Muraven and Baumeister, 2000). According to this model, the ability to exhibit self-control relies on a limited resource, or self-control strength, and all different self-control operations draw on that same resource. Their study (Muraven et al., 1998) showed that participants' performance was impaired in a self-control task that followed an initial one. In addition, the impairment was found even if the two tasks were completely different in context. The model's view of self-control resembles a muscle whose short-term ability decreases after exertion, but at the same time repeated exercise strengthens it in the long-run. In another study (Muraven et al., 1999), a group of students was asked to regularly perform some easy self-control tasks for two weeks. These participants showed significant improvements on self-control compared with participants who did not practice self-control.

Inspired by the study of Muraven et al. (1999), we tried to investigate whether a low-conflict matrix could serve as an exercise that would enhance the system's overall performance in a more demanding task. More specifically, both a low-conflict (Payoff Matrix (i) in Figure 5.1) and a very intense-conflict (Payoff Matrix (iii) in Figure 5.1) payoff structures were used in the simulation such that they interchanged each other every ten rounds of the game. The agents' competition started with the low-conflict matrix as the payoff matrix of the game and was kept constant for ten consecutive rounds, by which

time it was replaced by the intense conflict-matrix for another ten rounds. The game's procedure continued in the same manner by changing between the two matrix structures where the low-conflict matrix served as a self-control exercise for the demanding intense-conflict task that followed.

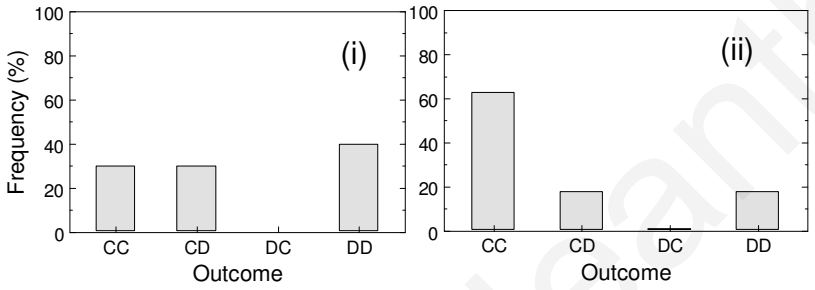

Figure 5.6: Exercising self-control similar to a body muscle helps resolving intense internal conflicts.

Average outcomes after 200 rounds of the IPD with (i) constant intense-conflict payoff structure and (ii) when intense-conflict payoff structure interchanged with a low-conflict payoff structure. Practicing self-control on the low-conflict structure improved the ability of the simulated individual to resolve a more intense internal conflict through self-control. The increase in the CC outcome as well as the decrease in the DD outcome from case (i) to (ii) is statistically significant using a one-tailed z-test at 95% confidence interval.

Results (Figure 5.6) demonstrate a significant increase in the cooperative outcome compared to the simulation where only the intense-conflict matrix was used and a decrease in the respective CD and DD outcomes. The CC outcome now resulted 63% of the times compared to the previous 30% whereas both the CD and DD outcomes dropped from 30% and 40% respectively to just 18%. In the case where only the intense-conflict matrix was used, the large absolute values of the Temptation and Sucker payoff induced a premature sensitization of the respective agent towards the immediate payoff which is

attained through defection. Learning in that case worked against mutual cooperation as it induced great synaptic changes that promoted defection from the column agent very early on in the learning process. As a result the system could not effectively learn to exploit long-term payoff and resulted in high percentages of the CD and DD outcomes. On the other hand, the intervention of the low-conflict matrix during the simulation enabled the system to discover and 'appreciate' the value of mutual cooperation with respect to a greater long-term payoff. Learning took place in a smoother and more effective way allowing the system to explore the different strategies and choose the one that will benefit it the most in the long run. Our findings not only agree with the theory that perceives self-control as a quality that can be exercised (Muraven et al., 1999), but also provide a possible interpretation for it. Practising self-control builds up an appreciation for the long-term, accumulated payoff provided by self-control. In addition, this appreciation is enabled by learning that takes place more effectively in an easy self-control task where it is easier to resist temptation and prevent sensitisation by the immediate payoff. Once the appreciation is built, then it is easier to employ self-control in more difficult tasks, even outside the context that has been practiced.

Chapter 6

Conclusions and Future Work

6.1 Overview of the Computational Model of Internal Conflict

The current research thesis proposes a novel computational model of internal conflict. The model integrates knowledge from such diverse areas such as psychology, game theory, neuroscience and computational neuroscience, while it is applied on a multiagent reinforcement learning task (IPD) with the aim to promote understanding in a psychological problem. More specifically, the system models internal conflict as a strategic interaction between internal rational subagents of the brain (Kavka, 1991), which can be optimally resolved through self-control behaviour. We implemented this simple, elegant and powerful game theoretical view of internal conflict in a neurobiologically relevant computational model of spiking neural networks that learn through biologically plausible learning algorithms. The model does not intend to reproduce in detail the actual brain regions that are involved in internal conflicts. Too much information is missing with respect to the precise identity and function of the regions that enable and influence this highly complex state of mind. Even if these regions were readily available to us and their function was understood in detail, it would probably take more than a lifetime to model in detail the actual brain in terms of its structure and function. In addition, the existence of a computational model in that case would have

been somehow redundant, as knowing in detail the actual brain would question the necessity for the existence of a computational model unless of course it would be for the purposes of creating artificial life. Whether this is possible or how distant this scenario lies, is a controversial issue out of the scope of this discussion. The current fact is that internal conflict and self-control behaviour are poorly understood. Given that, we believe that a high-level computational model that implements a game theoretical view of internal conflict, rather than an actual detailed reproduction of the brain areas involved, helps us to acquire a better understanding of the big picture of internal conflict (and how it can be resolved), which is currently obscure. However, our computational model does not disregard important experimental findings with respect to intertemporal choice (e.g., McClure et al., 2004), which is the context in which internal conflict generally resides and self-control behaviour could be exhibited. These identified regions and their functions were integrated in our game theoretical computational model by providing a plausible view of how the competing agents and these brain regions might relate. More specifically, we propose that each of these internal agents has access to, or "entail", both regions that are involved in the valuations of delayed and immediately available payoffs (Figure 2.4). Competition in the model exists between higher (delayed reward valuations) and lower processes (immediate reward valuations) of the brain within each subagent, but also between subagents who represent distinct evaluation criteria (other than temporal). Their actions whether to choose the long term or the immediate reward are determined by the relative activation of these regions which is also consistent with experimental findings. It is noted that although it is a simple and abstract approach to modelling internal conflict, this is a novel view and the only view that allows internal conflict to be properly modelled by the IPD when including these brain regions in the interpretation. Other work

that also used the IPD to model internal conflict (Banfield, 2006; Rachlin, 2000), regarded that each agent should comprise either of the two brain regions, which is problematic as it suggests that the limbic system should consider long term rewards. In order to model internal conflict with the IPD it is required that both agents should consider long term and immediate rewards, therefore each agent should have access to both of these brain regions.

Moreover, we regard that computational model is highly applicable to many real life situations due to the particular form of internal conflict that implements. We did not simply model internal conflict in the general context of intertemporal choice, even though it is highly present in our work since the agents are required to choose between long term superior rewards which are attained through self-control behaviour and immediate but smaller, in the long-run, rewards. In addition to the general context of intertemporal choice, our computational system models a particular form of internal conflict which resides in situations where a number of options can be evaluated along different dimensions of evaluation (as in the student's example in Section 2.2). This particular form of internal conflict is quite common, experienced by all people in their everyday life as people are asked to decide between available options that can be evaluated along divergent criteria, rather than between immediate and long term payoffs as such. Of course some people could consciously decide on the options by considering long term *versus* immediate payoffs and this makes our model even more relevant. In general, our model is applicable to an unlimited number of real life situations because this particular game theoretical view of internal conflict models every situation of internal conflict as long as there are two dimensions of evaluation that induce different preference orderings (again as in the student's example in Section 2.2). Moreover, the validity of the results

obtained by the model does not rely on the actual existence of internal subagents. Even if real subagents do not exist, thinking about and simulating internal value conflicts (in such evaluation-subordering situations) as if they were represented by rational subagents, can help us understand the dynamics and the variables that determine how these internal value conflicts are resolved. As Kavka (1991) put it "I do not claim that we are in fact composed of multiple distinct selves, each of which forms an integrated unit over time and has different dispositions or values from the other selves of the same individual. But I do think we can learn something about the structure and significance of internal value conflicts by treating different value-dimensions as represented by distinct rational subagents." Overall, for all the reasons given above as well as for the results presented in this thesis and overviewed below (Section 6.2), we believe that the most valuable contribution of this thesis work is the computational model itself.

6.2 Overview of Results Obtained by the Computational Model of Internal Conflict

What follows is an overview of the results obtained by the computational model of internal conflict which was developed in the course of this thesis work.

Overall, results obtained in Chapter 4 by both implemented algorithms, reinforcement of stochastic synaptic transmission (Seung, 2003) and reward-modulated STDP with eligibility trace (Florian, 2007), show that self-control behaviour can be learned. The system learned to establish a strong self-controlled behaviour, reflected by a strong *CC* outcome which was consistently and persistently obtained during the IPD. The simulated internal agents achieved through learning to postpone immediate gratification in favour of a superior outcome in the long run. Learning effectively regulated the activation of the output units such that to produce the self-control outcome.

Our computational results support the analytical work of Livnat and Pippenger (2006), which proposed that an optimal brain can be composed of "selfish" conflicting agents. This is the case, because the competing subagents in our system managed to learn to adopt the optimum strategy for themselves and the collective, maximising thus returns for the brain. Therefore our results further suggest that self-control behaviour is a learned maximising strategy employed by an optimal brain in the presence of conflicting value systems.

In addition, results showed that recurrent experience helped the system to learn self-control behaviour through correct associations between the available actions and the outcomes in terms of gains in reward both in the short term and in the long run. Note that the difficulty in learning such associations in our system as in real life, lies in the fact that the same action may result in opposite outcomes. Moreover, recurrent experience allowed the system to acquire an appreciation of the accumulated reward, which was obtained by exercising self-control behaviour consistently. Appreciation in our system is enabled by learning through the consistent administration of positive reinforcement signals and is reflected by the activation of the specific output units such as to enable persistent self-control behaviour. Therefore, in addition to experimental findings that showed a greater fronto-parietal activity in subjects who chose long-term rewards in intertemporal choice tasks (e.g., McClure et al., 2004), our results suggest that people who consistently and efficiently practise self-control in their everyday lives should have a greater respective neural activity induced by the appreciation towards the overall accumulated reward resulted from consistent practice of self-control behaviour.

Simulations revealed that the reward-correlated memory of the competing agents, facilitated by the variable of eligibility trace and its time constant, is also important in the

process of learning self-control behaviour since the memory is employed by the agents in choosing their actions such as to maximise their individual reward. Agents with stronger memory induced the best performance to the system since they learned to cooperate fast and therefore the system exhibited self-control behaviour quite early in the learning process. Agents with very low memory were not always able to learn how to cooperate and in the cases they did, it took longer. In addition, such agents demonstrated high percentages of mutual defection. Finally, in the simulation where one agent had strong memory and the other had weak, the performance of the system was somewhere in between. In addition the short-memory agent exploited the strong-memory agent by unilateral defection until the latter also changed its strategy to defection. Therefore, the short-memory agent accumulated a higher total individual payoff by pursuing in a myopic manner immediate gratification through the Temptation payoff. If we accept that some internal subagents are more myopic than others (distinguished by the time they usually wait in order to collect the rewards), then these results suggest that myopic subagents (e.g., satisfy-hunger or taste agent) are more likely to exploit subagents that are less myopic and are willing to compromise (e.g., physical fitness agent). This might provide a psychological plausible picture for example of why we sometimes excess on sweets while dieting does not begin on Mondays.

Overall, results with respect to the effect of the eligibility trace time constant showed that strong reward-directed memory is important for the attainment of self-control behaviour. As in our computational model, this kind of memory might not concern the actual actions performed by the individual, but be implicitly present in the organism for optimisation purposes. However its reward-driven nature enables maximisation by the

individual through the performed actions as if the individual had an explicit memory on these actions.

In Chapter 5, simulations revealed that learning and recurrent experience are not sufficient for the attainment of self-control behaviour when internal conflict is intense. While the gain from obtaining maximum immediate gratification (Temptation payoff), and the loss from obtaining the least immediate gratification (Sucker's payoff) were getting higher, and therefore intensifying the conflict, the cooperative outcome was getting lower. In contrast, mutual defection was getting higher. The respective payoff structure promoted mutual defection whether only one of the agents had this particular payoff structure, or both. There was no significant difference in the cooperative outcome between the two cases. The difference was only in the percentage of mutual defection, which was higher in case where both agents had the respective structure. In cases where only one agent had this particular structure, results reveal increased percentages of unilateral defection from the same agent. In all cases, a relatively low immediate self-control outcome (CC) restricted the agents from building a quick appreciation for the long-term accumulated payoff obtained from steady mutual cooperation and thus made the possibility of mutual cooperation even harder to achieve.

These payoff structures characterise situations of internal conflict where, as shown by the simulations, self-control behaviour is very hard to achieve. More specifically, in cases where the structure of internal values along a given value-dimension is such that the most preferred present outcome yields much more than the present self-control outcome (e.g., smoking as opposed to not smoking for a single day) then it is very hard to resist the most preferred outcome for the sake of a long term reward. In that case it is likely that the most preferred outcome will be chosen and probably be collected (shown by increased

unilateral defection). On the other hand, if such a value structure exists in both dimensions of evaluation, then the outcome will probably not satisfy any of the two dimensions of evaluation as the individual will end up with a suboptimal outcome (shown by the increased mutual defection). The simulations help us understand the dynamics that influence the resolution of internal conflicts and can be used so as to avoid engaging ourselves in such situations where self-control is hard to achieve.

In a particular, in a simulation of intense conflict, the computational system exhibited sensitization similar to the one induced on the real biological neural system by addictive substances, in the sense that it persisted in choosing a respective action even after the "likeness" about the respective outcome changed.

Finally, the last simulation of Chapter 5 supports to the theory that perceives self-control strength as a muscle that can be exercised (Muraven et al., 1999).

6.3 Contributions

The presented thesis work contributed both to the problem of understanding internal conflict and self-control behaviour as well as to the more general area of spiking neural networks and reinforcement learning. We present below a list of the contributions resulted from this work.

- The computational model of internal conflict: for its novelty and applicability as well for behaving consistently in a game theoretical framework. A version of the computational model appeared in Christodoulou et al. (2010) as well as in Cleanthous and Christodoulou (2010, 2009a, 2009b).
- The results obtained by the system as presented in Section 6.2. Particular contributions as extracted from the respective results are:
    - Self-control behaviour can be learned.
    - Strong reward-directed memory is important for the attainment of self-control behaviour.
    - Learning and recurrent experience is not sufficient for the attainment of self-control behaviour when internal conflict is intense.
    - The research identified several internal value structures that promote or hinder the attainment of self-control behaviour.
    - The study confirms through computational modelling that an optimal brain can be composed of conflicting "selfish" agents, as suggested in a relevant theoretical study (Livnat and Pippenger, 2006), since the artificial neuronal system implemented the optimum in a strategic interaction of "selfish" agents. Results further show that self-control behaviour is a learned maximising

strategy employed by an optimal brain in the presence of conflicting value systems.

- Results support to the theory that perceives self-control strength as a muscle that can be exercised (Muraven et al., 1999).

Some of these results have been published in Christodoulou et al. (2010), Cleanthous and Christodoulou (2010, 2009a, 2009b).

- To the best of our knowledge, the current thesis work applies for the first time spiking neural agents combined with biological plausible reinforcement learning in a highly demanding multiagent task. In particular, it evaluates the effectiveness of reward modulated STDP and reinforcement of stochastic synaptic transmission in the general-sum game of the IPD. Results showed that both investigated learning algorithms exhibited 'sophisticated intelligence' in a non-trivial task. The spiking agents showed a capacity for playing the game along the lines of game theory in a way that resembles the behaviour of real players. During most simulations, the networks managed to adapt to the challenges of the game and make decisions according to the other player's decisions in order to maximise their accumulated payoff. Most importantly, they "displayed intelligence" because when the game flow allowed for the Pareto optimum solution to be reached they "took advantage of the possibility" and settled to the solution by choosing cooperation for the rest of the game. These results have been accepted for publication (Vassiliades et al., 2010).
- In general, from our experiments in both studied tasks, the XOR and the IPD, findings suggest that the increased firing irregularity at high rates, which results from the introduction of the partial somatic reset mechanism at every LIF neuron of the spiking

neural networks, enhances learning. Results showed increased suppression of the output firing rate for input pattern {1, 1} in relation to the output firing rates for input patterns {0, 1} or {1, 0} in the XOR problem and the resulting accumulation of higher cooperative reward in the IPD task. More specifically, this high firing irregularity at high rates enhances reward-modulated STDP with eligibility trace. We believe that this is due to more accurate correlations between pre-synaptic and postsynaptic spike timings and reinforcement signals. If firing is regular, then it is possible for two identical spike pairs to be associated with opposite in sign reinforcement signals, confusing thus the direction of the plasticity for a given synapse. High firing irregularity prevents this unnecessary competition by weakening this possibility and thus preventing a possible corruption of the learning algorithm. Given this justification, then the result should not apply exclusively for the learning algorithm implemented in this thesis (Florian, 2007) work but also to other variants of reward-modulated STDP like in Izhikevich (2007), Faries and Fairhall (2007) and Legenstein, Pecevski and Maass (2008). These results have been accepted for publication (Christodoulou and Cleanthous, 2010a; 2010b).

- The current work extended the reinforcement learning algorithms with additional, opposite in sign global reinforcement signals that were concurrently administered along with the signals prescribed by the original algorithms. The administration of additional global reinforcement signals, which increased competition at the neuronal and synaptic level, proved both novel and necessary for the high performance of the algorithms. These results have been accepted for publication (Vassiliades et al., 2010).

- The successful application of the learning algorithms to the IPD required high values of eligibility trace time constants for both networks. It follows that the extent to which the reinforcement applies to events happened in the past, determines the success of the learning algorithms. Results showed that reinforcement should apply to events over a longer period of time as that agents with a "stronger memory" configuration achieved the best cooperative result, indicating the importance of reward-directed memory in effective MARL. These results have been accepted for publication (Vassiliades et al., 2010).

6.4 Future Work

The current research thesis aimed at providing some further understanding on the perplexing and highly complex behaviour of self-control and internal conflict through an abstract but at the same time biologically and psychologically relevant computational model of internal conflict. We believe that we contributed in the right direction given the aim of the thesis. However, several aspects concerning both internal conflict and the associated self-control behaviour need further investigation, explanation and understanding. We are confident that the present computational model can be a starting point for the development of a more sophisticated computational system that will integrate further knowledge from the related scientific areas such that to implement the problem in a more biologically and psychologically realistic way.

More specifically, an important upgrade to the system will be achieved when emotions are integrated into the decision process. In the current thesis emotions are implicitly involved by regarding that the values of the outcomes, as summarized by the payoff matrix of the game, are shaped partly because of the emotional states they elicit. A

real challenge is to investigate in more detail how exactly emotions are involved in the process of shaping these values and how they can affect the overall outcome of the conflict. In addition, integration of emotions could be achieved through extra signals that disturb the overall structure of the value systems during an ongoing simulation and thus enhance the dynamic elements of the system. A theory that might be considered for implementing is the somatic markers hypothesis (Damasio, 1996). According to the theory, when an individual faces a decision, each alternative elicits a bodily state – a somatic marker – that corresponds to an emotional reaction. These reactions are believed to influence decision-making even in the absence of conscious reasoning. For example, the somatic markers could be implemented in our system as an additional feedback that can differentiate the payoff matrix.

An equally important update will be obtained when additional structural and functional neurobiological realism is integrated into the system. At this stage, the most valuable enhancement would be a detailed incorporation of the signals to and from the simulated brain areas such that a more biologically realistic interplay between them is implemented.

As far as the existing computational model is concerned, more insight into the operation will be achieved by investigating the role of excitation and inhibition as well as their routes of action in the system, which leads to the observed behaviour. In addition, as suggested by one of the examiners, if one assumed immediate learning, it turns out that the system under our non-conventional training scheme behaves like a state machine, moving from state to state, e.g., CD or DC -> DD -> CC with CC as a stable fixed point. This interesting point should be further explored. Moreover, more experiments should be conducted using variable payoff matrices as well as it would be interesting to see whether

high firing irregularity enhances learning in other than STDP related learning algorithms such as in Seung's reinforcement of stochastic synaptic transmission (Seung, 2003).

Furthermore, the relation between self-control behaviour and subjective experience (or "consciousness'') could be investigated. According to Morsella et al. (2009) conflicts involving delay of gratification (Mischel et al., 1989), a self-control problem, lead to systematic changes in subjective experience. If we attempt to interpret our results though the theory developed by Morsella et al. (2009), then we could ask whether learning self-control behaviour has any impact on the perturbations in consciousness. In particular, one may wonder whether the transition from weak to strong self-control behaviour through learning could indicate a transition in the level consciousness. Our preliminary ideas on this issue have been presented in a conference (Christodoulou and Cleanthous, 2009).

References

Ainslie, G. (1975). Specious reward: a behavioral theory of impulsiveness and impulse control. Psychological Bulletin, 82, 463–496.

Ainslie, G. (2001). Breakdown of will. Cambridge University Press, Cambridge, UK.

Ainslie, G. and Haendel, V (1983). The motives of the will. In etiologic aspects of alcohol and drug abuse, ed. by Gottheil, E., Duley, K., Skoloda, T. and Waxman, M. Charles C Thomas. Springfield, IL. 119-140.

Arena, P., Fortuna, L., Frasca, M. and Patane, L. (2009). Learning anticipation via spiking networks: application to navigation control. IEEE Transactions on Neural Networks, 20(2), 202–216.

Ariely, D. (2002). Procrastination, deadlines and performance: self-control by precommitment. MIT Press, Cambridge, MA.

Axelrod, R. and Hamilton, W.D. (1981). The evolution of cooperation. Science, 211, 1390-1396.

Banfield, G. (2006). Simulation of self-control through precommitment behaviour in an evolutionary system. Ph.D Thesis, Birkbeck, University of London.

Barkley, R. A. (1997). ADHD and the nature of self-control. Guilford Press, New York.

Baxter, J., Bartlett, P. L. and Weaver, L. (2001). Experiments with infinite-horizon, policy-gradient estimation. Journal of Artificial Intelligence Research, 15, 351–381.

Bell, C.C., Han, V. Z., Sugawara, Y. and Grant, K. (1997). Synaptic plasticity in a cerebellum-like structure depends on temporal order. Nature, 387(6630), 278–281.

Bembenutty, H. (2009). Academic delay of gratification, self-regulation of learning, gender differences, and expectancy-value. Personality and Individual Differences, 46, 347-352.

Benzion, U., Rapoport, A. and Yagil, J. (1989). Discount rates inferred from decisions: An experimental study. Management Science, 35, 270-284.

Berridge, K.C. and Robinson, T.E. (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? Brain Research Reviews, 28, 309-369.

Berger, S.P., Hall, S., Mickalian, J.D., Reid, M.S., Crawford, C.A., Delucchi, K., Carr, K. and Hall, S. (1996). Haloperidol antagonism of cue-elicited cocaine craving. Lancet, 347, 504-508.

Bi, G.Q. and Poo, M.M. (1998). Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. Journal of Neuroscience, 18(24), 10464–10472.

Brauer, L.H. and DeWit, H. (1996). Subjective responses to d-ampetamine alone and after pimozide pretreatment in normal, healthy volunteers. Biological Psychiatry, 39, 26-32.

Brauer, L.H. and DeWit, H. (1997). High dose pimozide does not block amphetamine-induced euphoria in normal volunteers. Pharmagology, Biochemistry and Behaviour, 56, 265-272.

Brown, J.S. (1948). Gradients of approach and avoidance responses and their relation to level of motivation. Journal of Comparative and Physiological Psychology, 41, 450–465.

Bugmann, G., Christodoulou, C. and Taylor, J. (1997). Role of temporal integration and fluctuation detection in the highly irregular firing of a leaky integrator neuron model with partial reset. Neural Computation, 9, 985–1000.

Carver, C. S. and Scheier, M. F. (1982). Control theory: A useful conceptual framework for personality-social, clinical and health psychology. Psychological Bulletin, 92, 111-135.

Carver, C. S. and Scheier, M. F. (1998). On the self-regulation of behavior. Cambridge University Press, New York.

Christodoulou, C., Banfield, G. and Cleanthous, A. (2010). Self-control with spiking and non-spiking neural networks playing games. Journal of Physiology - Paris, 104, 108-117.

Christodoulou, C. and Bugmann, G. (2001). Coefficient of variation (CV) vs mean interspike interval (ISI) curves: what do they tell us about the brain. Neurocomputing, 38-40, 1141–1149.

Christodoulou, C., Bugmann, G. and Clarkson, T.G. (2002). A spiking neuron model: applications and learning. Neural Networks, 15 (7), 891–908.

Christodoulou, C. and Cleanthous, A. (2009). Modelling and resolving conscious conflict through learned self-control behaviour. Proceedings of the Conference Consciousness and its Measures. Limassol, Cyprus, November/December 2009, 27-28.

Christodoulou, C. and Cleanthous, A. (2010a). Does high firing irregularity enhance learning? Neural Computation (in press).

Christodoulou, C. and Cleanthous A. (2010b). Spiking neural networks with different reinforcement learning schemes in a multiagent setting. Chinese Journal of Physiology, 53(6), doi: 10.4077/CJP.2010.AMM030 (in press).

Cleanthous, A. and Christodoulou, C. (2009a). Is self-control a learned strategy employed by a reward maximizing brain? BMC Neuroscience, 10 (Suppl 1): P14.

Cleanthous, A. and Christodoulou, C. (2009b). On the psychology and modelling of self-control. In: Connectionist Models of Behaviour and Cognition II, Progress in Neural Processing, ed. by J. Mayor, N. Ruh, K. Blunkett. World Scientific, 18, 229-240.

Cleanthous, C. and Christodoulou, C. (2010). How dynamical changes in the payoff matrix of the Iterated Prisoner's Dilemma enhance the understanding of how to attain self-control behaviour. Proceedings of the 14th International Conference on Cognitive and Neural Systems, Boston, USA, May 2010, 67.

Cohen, J. D., Dunbar, K. and McClelland, J. (1990). On the control of automatic processes: a parallel distributed processing account of the stroop effect. Psychological Review, 97, 332-361.

Damasio, A. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. Philosophical Transactions of the Royal Society of London, 351, 1413-1420.

Dan, Y. and Poo, M.M. (1992). Hebbian depression of isolated neuromuscular synapses in vitro. Science, 256(5063), 1570–1573.

Dan, Y. and Poo, M.M. (2004). Spike timing-dependent plasticity of neural circuits. Neuron, 44, 23–30.

De Martino, B., Kumaran, D., Seymour, B. and Dolan, R.J. (2006). Frames, biases, and rational decision-making in the human brain. Science, 313, 684–687.

Faries, M.A. and Fairhall, A.L. (2007). Reinforcement learning with modulated spike-timing-dependent synaptic plasticity. Journal of Neurophysiology, 98, 3648-3665.

Florian, R.V. (2007). Reinforcement learning through modulation of spike-timing dependent synaptic plasticity. Neural Computation, 19, 1468-1502.

Fischman, M.W. (1989). Relationship between self-reported drug effects and their reinforcement effects: studies with stimulant drugs. NIDA Research Monographs, 92, 211-230.

Fischman, M. W. and Foltin, R.W. (1992). Self-administration of cocaine by humans: a laboratory perspective. In: Cocaine scientific and social dimensions, ed. by Bock G.R. and Whelan, J. CIBA Foundation Symposium No. 166. Wiley, Chichester, UK. 165-180.

Fodor, J.A. (1983). The modularity of mind. MIT Press, Cambridge, MA.

Frederick, S., Loewenstein, G. and O'Donoghue, T. (2002). Time discounting and time preference: a critical review. Journal of Economic Literature, 40, 351–401.

Frederick, S., Loewenstein, G. and O'Donoghue, T. (2003). Time discounting and time preference. In: Time and decision, ed. by Loewenstein, G., Read, D. and Baumeister, R. Russell Sage, New York. 13-86.

Fudenberg, D. and Tirole, J. (1991). Game Theory. MIT Press, Cambridge, MA.

Gailliot, M.T. and Baumeister, R.F. (2007). The physiology of willpower: Linking blood glucose to self-control. Personality and Social Psychology Review, 11(4), 303–27.

Gossen, H.H. (1854) Die entwicklung der gesetze des menschlichen verkehrs und der daraus fließenden regeln für menschliches handeln, Vieweg, Braunschweig (transl. as: The Development of the Laws of Human Intercourse and the Consequent Rules of Human Action Derived Therefrom, MIT Press, Cambridge, MA, 1983).

Green, L., Fry, A.F. and Myerson, J. (1994). Discounting of delayed rewards: a lifespan comparison. Psychological Science, 5, 33-36.

Herrnstein, R. (1981). Self-control as response strength. In: Quantification of steady-state operant behavior, ed. by Bradshaw, C., Szabadi, E. and Lowe, C. Elsevier-North Holland, New York. 3-20.

Herrnstein, R. (1990). Rational choice: Necessary but not sufficient. American Psychologist, 45, 356-67.

Hodgkin, A.L. and Huxley, A.F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. Journal of Physiology, 117, 500–544.

Hu, J. and Wellman, M.P. (2003). Nash-Q learning for general sum stochastic games. Journal of Machine Learning Research, 4, 1039-1069.

Izhikevich, E.M. (2003). Simple model of spiking neurons. IEEE Transactions on Neural Networks, 14(6), 1569–1572.

Izhikevich, E.M. (2004). Which model to use for cortical spiking neurons? IEEE Transactions on Neural Networks, 15(5), 1063–1070.

Izhikevich, E.M. (2007). Solving the distal reward problem through linkage of STDP and dopamine signalling. Cerebral Cortex, 17, 2443-2452.

Jacobs, R.A. (1999). Computational studies of the development of functionally specialized neural modules. Trends in Cognitive Sciences, 3, 31-38.

Kable, J.W. and Glimcher, P.W. (2007). The neural correlates of subjective value during intertemporal choice. Natural Neuroscience, 10, 1625–1633.

Kahneman, D. and Tversky, A. (1979). Prospect theory: an analysis of decision under risk. Econometrica, 47, 263-291.

Kanfer, F. H. and Karoly, P. (1972). Self-control: A behavioristic excursion into the lion's den. Behavioral Therapy, 3, 398-416.

Kavka, G. (1991). Is individual choice less problematic than collective choice?. Economics and Philosophy, 7, 291-310.

Kirby, K. (1997). Bidding on the future: evidence against normative discounting of delayed rewards. Journal of Experimental Psychology, 126, 54-70.

Kirby, K. and Herrnstein, R. (1995). Preference reversals due to myopic discounting of delayed reward. Psychological Science, 6, 83-89.

Kirby, K. and Marakovic, N. (1995). Modeling myopic decisions: evidence for hyperbolic delay discounting with subjects and amounts. Organizational Behavior and Human Decision Processes, 64, 22-30.

Klopf, A.H. (1982). The hedonistic neuron: a theory of memory, learning and intelligence. Hemisphere Publishing Cooperation, Washington, D.C.

Lamp, R.J., Preston, K.L., Schindler, C.W., Meisch, R.A., Davis, F., Katz, J.L., Henningfield, J.E. and Goldberg, S.R. (1991). The reinforcing and subjective effects of morphine in post-addicts: a dose-response study. Journal of Pharmagology and Experimental Therapeutics, 259, 1165-1173.

Lansky, P. and Musila, M. (1991). Variable initial depolarization in Stein's neuronal model with synaptic reversal potentials. Biological Cybernetics, 64, 285–291.

Lapicque, L. (1907). Recherches quantitatives sur lexcitation electrique des nerfs traitee comme une polarisation. Journal de Physiologie et Pathologie Generale, 9, 620–635.

Legenstein, R., Pecevski, D. and Maass, W. (2008). A learning theory for reward-modulated spike-timing-dependent plasticity with application to biofeedback. PLoS Computational Biology, 4(10), e1000180.

Lin, J.K., Pawelzik, K., Ernst, U. and Sejnowski, T.J. (1998). Irregular synchronous activity in stochastically-coupled networks of integrate and-fire neurons. Network: Computation in Neural Systems, 9(3), 333–344.

Littman, M. (1994). Markov games as a framework for multi-agent reinforcement learning. In: Proceedings of the Eleventh International Conference on Machine Learning. San Francisco. Morgan Kaufmann. 157-163.

Littman, M. (2001). Friend-or-foe Q-learning in general-sum games. In: Proceedins of the Eighteenth International Conference on Machine Learning, ed. by Brodley, C., Danyluk, A. Morgan Kaufmann. 322–328.

Livnat, A. and Pippenger, N. (2006). An optimal brain can be composed of conflicting agents. Proceedings of the National Academy of Sciences of the United States of America. 103, 3198-3202.

Loewenstein, G. (1987). Anticipation and the valuation of delayed consumption. The Economic Journal, 97, 666-84.

Loewenstein, G. and Prelec, D. (1992). Anomalies in intertemporal choice: evidence and an interpretation. The Quarterly Journal of Economics, 107, 573-597.

MacGregor, R.J. (1987). Neural and brain modeling. Academic Press, San Diego, CA.

MacGregor, R.J and Oliver, R.M. (1974). A model for repetitive firing in neurons. Kybernetik, 16(1), 53–64.

Markram, H., Lübke, J., Frotscher, M. and Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. Science, 275(5297), 213–215.

Mazur, J. (1987). An adjustment procedure for studying delayed reinforcement. In: The effect of delay and intervening events on reinforcement value, ed. by Commons, M., Mazur, J., Nevin, J. and Raiffa, H. Hillsdale, Lawrence Erlbaum, NJ. 55-73.

McClure, S.M., Ericson, K.M., Laibson, D.I., Loewenstein, G. and Cohen, J.D. (2007). Time discounting for primary rewards. The Journal of Neuroscience, 27, 5796–5804.

McClure, S.M., Laibson, D.I., Loewenstein, G. and Cohen, J.D. (2004a). Separate neural systems value immediate and delayed monetary rewards. Science 306, 503–507.

McClure, S.M., Li, J., Tomlin, D., Cypert, K.S., Montague, L.M. and Montague, P.R. (2004b). Neural correlates of behavioral preference for culturally familiar drinks. Neuron 44, 379–387.

Metcalfe, J. and Mischel, W. (1999). A hot/cool-system analysis of delay of gratification: dynamics of willpower. Psychological Review, 106, 3–19.

Miller, N. E. (1944). Experimental studies of conflict. In: Personality and behavior disorders, ed. by Hunt, J. Ronald Press. 431-465.

Minsky, M. (1985). The society of mind. Simon and Schuster, New York, NY.

Mischel, W. (1996). From good intentions to willpower. The psychology of action: linking cognition and motivation to behavior, ed. by Gollwitzer, P.M. and Bargh, J.A. Guilford Press, New York. 197-218.

Mischel, W., Shoda, Y. and Rodriguez, M. (1989). Delay of gratification in children. Science, 244, 933–938.

Morsella, E., Gray, J. R., Krieger, S. C. and Bargh, J. A. (2009). The essence of conscious conflict: subjective effects of sustaining incompatible intentions. Emotion, 9, 717-728.

Muraven, M. and Baumeister, R. F. (2000). Self-Regulation and deplation of a limited sources: Does self-control resemble a muscle? Psychological Bulletin, 126(2), 247-259.

Muraven, M., Baumeister, R. F. and Tice, D. M. (1999). Longitudinal improvement of self-regulation through practice: building self-control strength through repeated exercise. Journal of Social Psychology, 139, 446-457.

Muraven, M., Tice, D. M. and Baumeister, R. F. (1998). Self-control as a limited resource: regulatory depletion patterns. Journal of Personality and Social Psychology, 74, 774-789.

Nash, J. (1950). Equilibrium points in n-person games. In: Proceedings of the National Academy of Sciences of the United States of America. 36, 48–49.

Nisbett, R.E. and Wilson, T.D. (1977). Telling more than we can know: verbal reports on mental processes. Psychological Review, 84, 231-259.

Ohuoha, D.C., Maxwell, J.A., Thomson, L.E., Cadet, J.L. and Rothman, R.B. (1997). Effect of dopamine receptor antagonists on cocaine subjective effects: a naturalistic case study. Journal of Substance Abuse Treatment, 14, 249-258.

O'Reilly, R.C. and Munakata, Y. (2000). Computational explorations in cognitive neuroscience: understanding the mind by simulating the brain. MIT Press Cambridge, MA.

Pareto, V. (1906). Manuale di economia politica. Societa Editrice, Milan.

Pender, J. (1996). Discount rates and credit markets: theory and evidence from rural India. Journal of Development Economics, 50, 257-96.

Pfister, J., Toyoizumi, T., Barber, D. and Gerstner, W. (2006). Optimal spike-timing dependent plasticity for precise action potential firing in supervised learning. Neural Computation, 18, 1318–1348.

Potjans, W., Morrison, A. and Diesmann, M. (2009). A spiking neural network model of an actor-critic learning agent. Neural Computation, 21, 301-339.

Pychyl, T. (2009). URL: www.psychologytoday.com/blog/don't delay/200903/academic-delay-gratification-motivation-and-self-regulated-learning-strategie (accessed March 27, 2009).

Rachlin, H. (1995). Self-Control: beyond commitment. Behavioural and Brain Sciences 18, 109-59.

Rachlin, H. (2000). The science of self-control. Harvard University Press, Cambridge, MA.

Rappoport, A. and Chammah, A. M. (1965). Prisoner's dilemma. University of Michigan Press, Ann Arbor, MI.

Richmond, B.J., Liu, Z. and Shidara, M. (2003). Predicting future rewards. Science, 31, 179-180.

Robinson, T.E. and Berridge. K.C. (1993). The neural basis of drug craving: an incentive-sensitization theory of addiction. Brain Research Reviews, 18, 247-291.

Robinson, T.E. and Berridge. K.C. (2000). The psychology and neurobiology of addiction: An incentive-sensitization view. Addiction, 95 (Suppl. 2), S91-S117.

Samuelson, P. (1937). A note on measurement of utility. Review of Economic Studies, 4, 155-161.

Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E. and Cohen, J.D. (2003). The neural basis of economic decision-making in the ultimatum game. Science, 300, 1755–1758.

Sanfey, A.G., Loewenstein, G., McClure, S.M. and Cohen, J.D. (2006). Neuroeconomics: cross-currents in research on decision-making. Trends in Cognitive Science, 10, 108–116.

Scheier, M.R. and Carver, C.S. (1988). A model of behavioral self-regulation: Translating intention into action. In: Advances in Experimental Social Psychology, ed. by L. Berkowitz. Academic Press, San Diego, CA. 21, 303-339.

Seung, H. S. (2003). Learning in spiking neural networks by reinforcement of synaptic transmission. Neuron, 40, 1063-1073.

Singh, S., Barto, A.G. and Chentanez, N. (2005). Intrinsically motivated reinforcement learning. In: Advances in Neural Information Processing Systems 17, ed. by Saul, L.K., Weiss, Y. and Bottou, L. MIT Press, Cambridge, MA. 1281–1288.

Smith, A. (1759). The theory of moral sentiments. A. Miller, A. Kincaid, and J. Bell, London.

Smolensky, P. (1988). Putting together connectionism. Behavioral and Brain Sciences, 11, 59-70.

Softky, W. and Koch, C. (1992). Cortical cells should fire regularly, but do not. Neural Computation, 4, 643–646.

Softky, W. and Koch, C. (1993). The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. Journal of Neuroscience, 13, 334–350.

Stein, R.B. (1967). Some models of neuronal variability. Biophysical Journal, 7(1), 37–68.

Sutton, R.S. (1998). Learning to predict by the method of temporal differences. Machine Learning, 3, 9-44.

Sutton, R. S. and Barto, A. G. (1998). Reinforcement learning: an introduction. MIT Press, Cambridge, MA.

Swiercz, W., Cios, K.J., Staley, K., Kurgan, L., Accurso, F. and Sagel, S. (2006). A new synaptic plasticity rule for networks of spiking neurons. IEEE Transactions on Neural Networks, 17(1), 94–105.

Tanaka, S.C., Doya, K., Okada, G., Ueda, K., Okamoto, Y. and Yamawaki, S. (2004). Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. Nature Neuroscience, 7, 887–893.

Thaler, R. (1981). Some empirical evidence on dynamic inconsistency. Economics Letters, 8, 201-207.

Thaler, R. and Shefrin H.M. (1981). An Economic Theory of Self-control. The Journal of Political Economy, 89(2), 392-406.

Tinbergen, N. (1952). "Derived" activities; their causation, biological significance, origin, and emancipation during evolution. The Quarterly Review of Biology, 27, 1-32.

Vasilaki, E., Fremaux, N., Urbanczik, R., Senn,W. and Gerstner,W. (2009). Spike-based reinforcement learning in continuous state and action space: When policy gradient methods fail. PLoS Computational Biology, 5(12), e1000586.

Vassiliades, V., Cleanthous, A. and Christodoulou, C. (2010). Multiagent reinforcement learning: Spiking and non-spiking agents in the Iterated Prisoner's Dilemma. IEEE Transactions on Neural Networks (accepted, in press).

Von Neumann, J. and Morgenstern, O. (1947). Theory of games and economic behavior. Princeton University Press, Princeton, NJ.

Xie, X. and Seung, H.S. (2004). Learning in neural networks by reinforcement of irregular spiking. Psychological Review, E69, 41909.

Xu, L., Liang, Z.Y., Wang, K., Li, S. and Jiang, T. (2009). Neural mechanism of intertemporal choice: from discounting future gains to future losses. Brain Research, 1261, 65-74.

Widrow B., Gupta, N. K. and Maitra, S. (1973). Punish/reward: learning with a critic in adaptive threshold systems. IEEE Transactions Systems, Man and Cybernetics, 3, 455-465.

Wittmann, M., Leland, D.S. and Paulus, M.P. (2007). Time and decision making: differential contribution of the posterior insular cortex and the striatum during a delay discounting task. Experimental Brain Research, 179, 643–653.

Appendix: List of Publications

Articles in Refereed Archival Journals

Christodoulou, C., Banfield, G. and Cleanthous, A. (2010). Self-control with spiking and non-spiking neural networks playing games. Journal of Physiology - Paris, 104, 108-117.

Vassiliades, V., Cleanthous, A. and Christodoulou, C. (2010). Multiagent reinforcement learning: Spiking and non-spiking agents in the Iterated Prisoner's Dilemma. IEEE Transactions on Neural Networks (accepted, in press).

Christodoulou, C. and Cleanthous, A. (2010). Does high firing irregularity enhance learning? Neural Computation (accepted, in press).

Christodoulou, C. and Cleanthous A. (2010). Spiking neural networks with different reinforcement learning schemes in a multiagent setting. Chinese Journal of Physiology, 53(6), doi: 10.4077/CJP.2010.AMM030 (in press).

Cleanthous, A. and Christodoulou, C. (2009). Is self-control a learned strategy employed by a reward maximizing brain? BMC Neuroscience, 10 (Suppl 1): P14.

Refereed Articles in Book Series and Compiled Volumes

Vassiliades, V., Cleanthous, A. and Christodoulou, C. (2009). Multiagent reinforcement learning with spiking and non spiking agents in the iterated prisoner's dilemma. In: Artificial Neural Networks - ICANN 2009, Lecture Notes in Computer Science, ed. by C. Alippi, M. Polycarpou, C. Panayiotou, G. Ellinas, Springer, 5768, 737-746.

Cleanthous, A. and Christodoulou, C. (2009). On the psychology and modelling of self-control. In: Connectionist Models of Behaviour and Cognition II, Progress in Neural Processing, ed. by J. Mayor, N. Ruh, K. Blunkett. World Scientific, 18, 229-240.

Refereed Abstracts

Christodoulou, C. and Cleanthous, A. (2010). High firing irregularity enhances learning. Proceedings of the 9th International Workshop on Neuronal Coding, Limassol, Cyprus, October/November 2010 (in press).

Cleanthous, C. and Christodoulou, C. (2010). How dynamical changes in the payoff matrix of the Iterated Prisoner's Dilemma enhance the understanding of how to attain self-control behaviour. Proceedings of the 14th International Conference on Cognitive and Neural Systems, Boston, USA, May 2010, 67.

Christodoulou, C. and Cleanthous, A. (2009). Modelling and resolving conscious conflict through learned self-control behaviour. Proceedings of the Conference Consciousness and its Measures. Limassol, Cyprus, November/December 2009, 27-28.

Vassiliades, V., Cleanthous, A. and Christodoulou, C. (2009). Multiagent Reinforcement Learning: Spiking and Non- spiking Neural Network Agents. Proceedings of the 2nd Cyprus Workshop on Signal Processing and Informatics, Nicosia, Cyprus, July 2009, 16.

Christodoulou, C. and Cleanthous, A. (2009). Modelling Self-Control Behaviour with Spiking Neural Networks in a Multiagent Reinforcement Learning Framework. Proceedings of the 2nd Cyprus Workshop on Signal Processing and Informatics, Nicosia, Cyprus, July 2009, 18.

Christodoulou, C. and Cleanthous, A. (2009). Spiking Neural Networks with Different Reinforcement Learning Schemes in a Multiagent Setting. Proceedings of the 8th International Workshop on Neuronal Coding, Tainan, Taiwan, May 2009, 57 59.

Christodoulou, C. and Cleanthous, A. (2008). On the psychology and modelling of self-control. Proceedings of the 11th Neural Computation and Psychology Workshop (Book of Abstracts), Oxford, UK, July 2008.

Cleanthous A. and Christodoulou, C. (2008). Can networks of leaky integrate-and-fire neurons with spike-based reinforcement learning play games? Workshop for Spiking Networks and Reinforcement Learning, Utah, USA. (available at: www. http://cosyne.org/wiki/Workshop_speaker _Aristodemos_Cleanthous).