Ατομική Διπλωματική Εργασία

ΘΕΤΙΚΑ ΚΑΙ ΑΡΝΗΤΙΚΑ ΣΥΝΑΙΣΘΗΜΑΤΑ ΣΕ ΕΝΑ ΥΠΟΛΟΓΙΣΤΙΚΟ ΜΟΝΤΕΛΟ ΑΥΤΟΕΛΕΓΧΟΥ

Ανδριανή Νικοδήμου

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μάιος 2020

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Positive and negative emotions in a computational model of self-control

Andriani Nikodemou

Επιβλέπων Καθηγητής
Χρίστος Χριστοδούλου

Η Ατομική Διπλωματική Εργασία υποβλήθηκε προς μερική εκπλήρωση των απαιτήσεων απόκτησης του πτυχίου Πληροφορικής του Τμήματος Πληροφορικής του Πανεπιστημίου Κύπρου

Μάιος 2020

Acknowledgments

I would like to especially thank my supervisor, Dr Chris Christodoulou for introducing me to this intriguing interdisciplinary research area and guiding me throughout all the research stages. His advices were priceless. I also want to extend my thanks to my fellow student Stelios Tymvios for the useful discussions we had regarding Reinforcement Learning. Finally, I am grateful to my family and close friends. I would have not made it without their heartfelt support.

Abstract

This thesis suggests an incorporation of emotions in a computational model of self-control. Self-control is defined as the dilemma between a smaller sooner reward (SS) and a large later reward (LL) (Rachlin, 2000), that is experienced as an internal conflict between the higher (prefrontal cortex) and lower (limbic system) parts of the brain. The nature of this conflict can be effectively simulated by the Prisoner's Dilemma game, in which the two agents represent the upper and lower parts of the brain. The two agents learn to cooperate and thus, achieve self-control by engaging in the Iterated Prisoner's Dilemma (IPD) through a Q-Learning model. The effects of the positive and negative emotions are simulated by changing the values of the IPD's payoff matrix T, R, P, S, which are the reinforcing signals the agents receive, according to a specified interval in rounds and a ratio, that is, a positive or negative value that is added to the payoff value. This method is based on Rolls (2012) definition of emotions that "are states elicited by […] instrumental reinforcers". By changing separately or in combination the values of the payoff matrix throughout the game, we would like to see whether and how the increment or decrement of positive and negative emotions affect the self-regulatory process. Therefore, we do not simulate the emotions *per se* but rather their presence, and their impact on self-control. Our simulations' results are in compliance with what the psychologists and neuroscientists suggest the effects of emotions on self-control are, an observation that makes our model cognitively adequate. Firstly, we observe that the increment of the R value which simulates the presence of positive emotions increased the levels of self-control compared to the results produced by using a constant payoff matrix. Moreover, the *absence* of negative emotions (decreased T, P, S) appeared to be as effective as the presence of positive ones when the interval of change was small. Conversely, the elimination of positive emotions during the presence of negative ones results in high levels of self-control failure. However, self-control was not achieved when the negative emotions were significantly decreased using large interval of change (increased P and S), which shows the necessity of guilt and fear, for example, when people exercise self-control. On the contrary, we managed to enhance self-control when the model was experiencing further negative emotions (decreased P and S) by using small ratios and intervals of change. We also found out that the agents would still achieve self-control, even in the presence of negative emotions caused by the internal cognitive

conflict (increased T), but only as long as it was not in combination with the presence of explicitly elicited negative emotions (increased T, decreased S).

Contents

Chapter 1

Introduction

1.1 Introduction

The purpose of this thesis is to investigate the effect of emotional arousal, and in particular the effects of positive and negative emotions on the self-control behavior. Self-control is defined by cognitive psychology as a dilemma of choosing between a large later (LL) reward versus a smaller sooner (SS) reward (Rachlin, 2000). For example, a person on diet has to resist to the immediate reward that is available at the current moment (SS reward), in order to achieve their long-term goal of losing weight (LL reward). These kinds of choices are so usual in everyday life; we have to sacrifice the current indulgence, if we want to obtain the future reward that we have defined to be a more desirable outcome. This kind of dilemmas takes place between the higher and lower parts of the brain which are associated with cognitive processed responses and intuitive-processed responses respectively. The higher part is identified as the brain's prefrontal cortex, whereas the lower part is the limbic system. The interaction between these two parts was confirmed to have the form of an internal conflict by neuroscientists who studied the activation of brain during the anticipation of long-term rewards and the exertion of self-control (Hare et al., 2009; McClure et al., 2004). To simulate the conflict we deploy the general-sum game of the Prisoner's Dilemma (Kavka, 1991), where each agent can insist on maximizing its own reward by defecting, but the best outcome for both of them is achieved only when the agents cooperate. The agents do not know what the other one will choose: to cooperate (C) or defect (D). All the combinations of states in which the agents can result in are CC, CD, DC, DD. The state of self-control is represented by the CC state.

Georgiou (2015) used reinforcing learning, and more specifically the Q-Learning algorithm, in order to examine whether the two Q-Learning agents who were representing the two parts of the brain and were engaging in the Iterated Prisoner's Dilemma (IPD), could learn to cooperate (CC state) and therefore, achieve self-control. Before Georgiou (2015), Banfield (2006) and Cleanthous (2010), had already modelled self-control behavior using neural networks and it was proven that the deployment of biological realistic spiking neural networks, was not necessary in order to successfully model self-control (Christodoulou et al., 2010).

However, all the above models that were mentioned did not incorporate the concept of emotions in any way. After all, self-control is a human process that inevitably is linked with emotions. The guilt one feels for example, when she has succumb to the temptation, or the desperation that an alcohol addict feels when s/he continuously fails in self-control, are negative emotions that their interplay with self-control cannot be overlooked. Is there a way to improve self-control when negative emotion is involved? Does someone who feels positive emotions, such as joy and confidence, fail to self-regulate in the same likelihood compared to someone that experiences negative emotions? Psychologists suggest that positive emotions promote self-control, whereas negative emotions impair it (Tice et al., 2004). However, the implications especially of negative emotions on self-control, turns out to be far more complicated. We are trained to feel guilt and fear when we set goals and we do not work towards them (Loewenstein & O'Donoghue, 2006). In this way, since we are feeling bad, we transfer the bad consequences of not resisting to the present, so we have motivation to resist. So eventually negative feelings might be necessary in self-control. Moreover, brain studies revealed not only the emotional arousal during the anticipated of reward (Knutson & Greer, 2008), but also that emotion processing limits the cognitive resources and thus interfere with rational decision-making (Cyders & Smith, 2008). Thus, negative *and* positive emotions can result in poor decisions during the self-control dilemma, when they are intense.

In order to investigate all the above scenarios, we needed to find a way to simulate the *effects* of the positive and negative emotions in self-control. We relied on the Rolls (2012) definition, which defines that emotions "are states elicited by rewards and punishers, that is, by instrumental reinforcers." The agents receive rewards and punishers during the IPD

according to the values of the payoff matrix, which are the Temptation (T), Reward (R), Punishment (P) and Sucker's (S) payoffs. These values are defined to *cause* agents' emotions. We classify them as negative and positive emotions, again according to Rolls (2012): "according to whether the reinforcer is positive or negative". Thus, the T and R which are positive values, are classified as producing positive emotional states, whereas P and S which are negative values, are considered to produce negative emotional states.

The purpose is to examine what happens to the self-control behavior when the presence of positive and negative emotions is increased or decreased. In order to simulate this effect, we will be changing the values of the payoff matrix throughout the rounds of the IPD. The change of positive reinforcers indicates the presence of positive emotions, and the change of negative reinforcers indicates the presence of negative emotions. The changes take place every few rounds that elapse, separately or in combination with others according to a ratio, that is, a positive or negative value that is added to the payoff value. The role of these two parameters is indicated by Rolls (2012) who stated that we are sensitive to not just the level of the reinforcer, but also to the change in magnitude and frequency that is received. In the next chapters, we are going to take a closer look to the implementation and design of the model, as well as the results that the simulations produced. We discuss whether and how our results are in alignment with what the Psychology and Neuroscience literature suggests on the implications of emotions on the self-control behavior.

1.1 Thesis outline

While Chapter 1 is an introduction, Chapter 2 presents the epistemological background of self-control behavior in psychological and neurobiological terms, the concept of emotions and the definition that helped us simulated the effects of emotions on self-control, as well as the findings that psychologists and neuroscientists already revealed about the impact of positive and negative emotions on the self-regulatory processes. We present the Prisoner's Dilemma game, and its variation, the Iterated Prisoner's Dilemma game, that was deployed to simulate the structure of self-control in the brain in our model. Lastly, we present the reinforcing learning algorithm of Q-Learning that was deployed by our model.

Chapter 3 concerns with the design and the implementation of the Q-Learning model and the two Q-Learning agents that represent the upper and down parts of the brain. The way that the effects of positive and negative emotions are simulated, is explained in detail. Chapter 4 presents the results of the simulations, for each case (positive and negative emotions), and discusses the most important findings. Finally, chapter 5 presents an overview, explains the findings, and correlates them to the existing literature of the relationship of emotions and self-control.

Chapter 2

Epistemological Background

2.1 Self-control

The study of understanding humans' self-control behavior originates from the ancient assumption that humans are the only animals that base their acts on *reason* and, thus they are rational (Kraut, 2018). However, if humans are indeed rational, then why do we need to exert self-control in the first place? That is, given that we know what is best for us *and* we are rational agents, then why are we not always choose the best for us? Modern cognitive science suggests that rationality comes in degrees defined by the distance of the thought or behavior from the *optimum* (K. E. Stanovich, 2012). In the same way, self-control can be thought as a capability humans have that contributes towards achieving the optimum behavior. After all, humans act irrationally especially in the presence of temptations, and as de Sousa (2007) puts it, "if human beings can indeed be described as

rational animals, it is precisely in virtue of the fact that humans, of all the animals, are the only ones capable of irrational thoughts and actions".

Self-control behavior has received special attention and has been extensively researched by psychologists, behavioral economists, and neuroscientists in the search of revealing its mechanisms. The goals of understanding the underpinnings of self-control vary according to the field of application. For example, the psychologists' role is to suggest methods for improving self-control, while marketers' goal is to find ways to bypass it.

The psychological definition of self-control is the ability to resist to the temptation and gain the immediate reward, and instead stay focused and wait for the long-term goal or reward (Reeve, 2014). Self-regulation is often distinguished from self-control, as the process of accomplishing a long-term goal through planning and applying the plan by exercising self-control. Self-control and self-regulation are abilities that everyone has to some extent, however, there are individual differences on physiological, emotional and cognitive levels, which define the self-regulatory processes (Calkins & Howse, 2004), (Berman et al., 2013). Self-control behavior though *can* be developed and improved. The Stanford marshmallow experiment was a study on the delay of gratification, where children were offered a choice between one marshmallow (small and immediate reward) and two marshmallows (large and later reward) if they waited for about fifteen minutes. The results of the original experiment showed that focusing attention on the future reward enhances the ability to delay gratification (Walter Mischel & Ebbesen, 1970).

What the Stanford marshmallow experiment also showed, is that, the children who delayed gratification tended to have better life outcomes (W. Mischel et al., 1989). In addition, self-regulation has also been identified as a necessary component for successfully functioning in the social world (Heatherton, 2011). Several more studies identified self-control as the foundation of human society and individual success in it (Duckworth & Seligman, 2005; Tangney et al., 2004). From dieting, commitment to a career path or managing interpersonal relationships, self-control is essential in order to resist to a doughnut, keep focus on practicing, or hold yourself when you are angry and ready to say nasty things. Therefore, understanding self-control and trying to improve it is of paramount importance.

2.1.1 The top-down model of self-control

Self-control behavior can be thought of as a dilemma of choosing between a large later (LL) reward versus a smaller sooner (SS) reward, and from the perspective that the LL reward results in a more desirable outcome in the long-term than in the short-term (Rachlin, 2000). Figure 2.1 illustrates the concept of the delay of gratification. A person that has the dilemma between the SS and the LL reward, has to exert self-control and not succumb to the temptation at time t2 where the value of the immediate reward becomes greater than the value of the future reward. At t1 the future reward has the greatest value, but as the temptation is getting closer, its value is increasing and as a result the preferences reverse. The discounting of delayed reward was also experimentally observed (Herrnstein, 1990; Kirby & Herrnstein, 1995). Consider for example, a student who faces the following dilemma. The student got an invitation for a party that takes place tonight, but he also has a quiz the next morning. Thus, he has to choose between the SS reward which is to accept the invitation, go to the party and have fun, and the LL reward which is to stay at home, focus on studying, perform well on the quiz the next day and ultimately have good grades at the end of the semester.

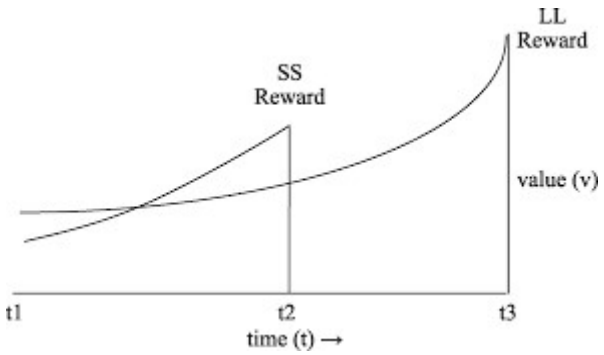

Figure 2.1: *Delay of gratification (based upon Rachlin, 1995). One has a choice between the SS reward at time t2 and the LL reward at time t3. Notice, that when the SS reward is available at time t2, its value is greater than the discounted LL reward. Exercising self-control means to not collect the immediate reward, but wait for the bigger one.*

Cognitive neuroscience suggests a model of self-control (Figure 2.2) which exactly depicts the internal process that an individual experiences, as our student in the aforementioned example. In this model, the state of the environment is perceived from

the higher center of the brain (prefrontal cortex) which is responsible for rational thinking, planning and control, and interacts with signals from the lower brain (limbic system), which is associated with emotion and is responsible for selecting an action. Finally, this internal conflict results in an action, which is rewarded or punished by stimuli from the external environment (Rachlin, 2000).

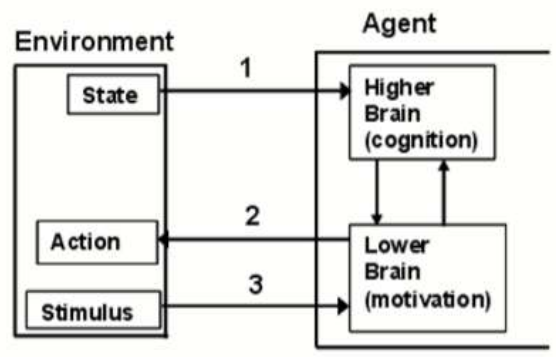

Figure 2.2: *A model of self-control behavior based upon Rachlin (2000). The higher brain perceives the state of the environment (Arrow 1). An action (or behavior) is generated in combination with the lower brain (Arrow 2). Stimulus reinforces the behavior by giving a reward or a punishment (Arrow 3).*

The neural basis of self-control has been well documented by neuroscientists who showed that there is increased activity in the prefrontal cortex (PFC) whenever one exerts self-control over temptation (Hare et al., 2009). Initially studies were concerned with the role of the prefrontal cortex in cognitive control (Miller & Cohen, 2001) and others proved its role in the top-down control over reflexive behavior (Curtis & D'Esposito, 2003), (Goldberg, 2002). The activation of the prefrontal cortex while anticipating long-term rewards and the activation of the limbic system with immediate rewards, are shown in brain-imaging studies (McClure et al., 2004) and thus, they confirm that an internal conflict takes place and the need of exercising self-control.

Interestingly enough, Kahneman (2011) introduces the concept of System 1 and System 2 that represent the intuitive and cognitive-processed responses respectively (Kahneman, 2011). This interpretation of the decision-making process is essentially consistent with the model of self-control in Figure 1.1; the higher part of the brain has the same characteristics with System 2, while the lower part has a similar function in respect to

System 1. According to Kahneman (2011), System 1 wants the fast reward, but System 2 knows that the longer it waits, the better. He also acknowledges how common is in our lives to experience "conflict between an automatic reaction and an intention to control it".

2.2 Prisoner's Dilemma

Kavka (1991) believed that the complexity of the collectivistic choices also exists in the nature of the individualistic choice. He used the prisoner's dilemma (PD) game to show how an individual experiences a value conflict. In the PD game (Rapoport et al., 1965), two agents have the choice to either cooperate for a good, common outcome, or to both selfishly defect. The goal is always to maximize their reward. The PD game is effectively described in the following scenario. The police arrests and interrogates two suspects — Alice and Bob— for committing a crime, in two separate rooms. They are told that if she confesses, that is *defect*, and her partner does not confess, she is going to be set free because she has helped the police, and Bob, the defected partner will be sentenced to 15 years in prison. The reverse scenario is also valid. However, in the case that both of them defect, they will be sentenced to 8 years in prison, not 15 because they have helped the police. There is also the scenario where both do not confess, that is *cooperate*, and end up in prison for less years (3 years each), because there is some evidence against them, but not strong enough, such as their confessions. Bob and Alice do not know what his/her partner will decide but know that he/she choose the best for him/herself therefore, to defect seems to be the best strategy. That is, Alice will not risk getting 15 years in prison by cooperating, in case Bob does not also cooperate. So, she will defect and the best scenario is that she will walk out free, in the case that Bob did not confess, and if the defected too, they will both end up 8 years in prison, and not 15 by herself. In his turn, Bob makes exactly the same analysis of the possible outcomes; he thinks that if Alice defects and he does not, then he will get the 15 years in prison. So, both of them, choose to defect. The strategy that is chosen by the players and from which nobody wants to deviate, is called a Nash equilibrium. Therefore, the outcomes of both defecting is considered the only Nash equilibrium (Nash, 1950) in the PD game. The dilemma then is that there is a better outcome than the mutual defection (8 years in prison) and that is, the

mutual cooperation (3 years), which however seems irrational from a self-interest perspective. Figure 2.3 shows all the possible scenarios.

		Bob	
		Cooperate	Defect
Alice	Cooperate	3, 3	15, 0
	Defect	0, 15	8, 8

Figure 2.3: *The number of years in prison for Alice and Bob according to the Prisoner's Dilemma game. Alice and Bob get 3 years in prison if they cooperate and do not confess and 8 years if they both defect and confess. If one of them defects and the other remains silent, the one who defected is set free, while his/her partner is sentenced to 15 years in prison.*

In the same way, since there is a need to simulate a value conflict, the higher and lower parts of the brain can be thought of two agents that have the choice to cooperate or defect. The four possible states are extracted by the two possible actions; the CC state is the result of mutual cooperation, the DD state is a result of mutual defection, and the CD or DC states are the result of one of the players cooperating and the other defecting. We define the CC state, as the goal of self-control behavior, and since it is the best outcome for both of the players. Below are the choices of the student who experiences a value conflict (have fun or have good grades), and how it resembles with the PD game (Christodoulou et al., 2010):

a. Go to the party, which corresponds to cooperation from the brain's lower part.
b. Stay home and study, which corresponds to cooperation from the brain's higher part.
c. Go to the party for some time and return home to study, which corresponds to mutual cooperation (CC state).
d. Do nothing, which corresponds to defection from both sides (DD state).

In the context of the game, values are assigned to the four different outcomes (a, b, c, d) which have to satisfy two rules. Firstly, the agent's reward (or payoff value) for defecting while the other is cooperating (Temptation) must be greater than the agents' payoff value for mutual cooperation (Reward). Secondly, Reward must be greater than the agents'

payoff if both defect (Punishment), which in turn is greater than the payoff of the agent that cooperated while the other defected (Sucker's payoff). That is, the payoffs must satisfy the equation (1). Figure 2.4 shows the payoff matrix of the Prisoner's Dilemma game.

$$\text{Temptation} > \text{Reward} > \text{Sucker's} > \text{Punishment} \qquad (1)$$

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

Figure 2.4: *The payoff matrix of the Prisoner's Dilemma game. Each player either cooperates or defects. When both cooperate, they both receive a Reward payoff (R), but when they both defect, they receive a Punishment payoff (P). When one of the players cooperate and the other defects, they receive the Sucker's payoff and Temptation payoffs, respectively.*

2.2.1 The Iterated Prisoner's Dilemma

The Iterated Prisoner's Dilemma (IPD), in which the agents play the game repetitively, is going to be used in the implementation. After all, mutual cooperation (CC state), which is the best outcome for both of the agents, is not obtained if the game is only played once. That is, in contrast with the IPD, the simple single-shot prisoner's dilemma game does not allow incorporating learning in the decision-making process. The IPD introduces a second rule that the game must satisfy which guarantees that if the profit of one agent increases, the profit of the other decreases.

$$2R > T + S \qquad (2)$$

The state that the two agents result at the end of each round of the IPD, represents the decision that the person took to solve his/her dilemma. If the state is the CC state, that means the agents cooperated and self-control was achieved. Each decision's round might be or not be about the same self-control task. After all, self-control is a behavior that can

be learnt by practicing it for one task, but once learnt, is demonstrated throughout all the individual's decisions and actions, which might be completely unrelated with the task that he/she was trained with.

Other general-sum games than the PD game like the Battle of the Sexes game (BoS) or the Rubinstein's Bargaining Game (RBG) (Rubinstein, 1982), could be considered to simulate the interaction between the higher and the lower brain. The RBG has fundamentally different structure than the PD game; the two players have to agree on how to split the reward, but if they do not agree, the reward decreases and a new offer takes place until they reach an agreement. Banfield (2006) simulated the RBG using Selective Bootstrap and Temporal Difference Reinforcement Learning for the respective RBG players, but despite that they seemed to have learnt (they would make less mistakes) it was realized that the nature of the game was not ideal. The BoS game is based on a payoff matrix but has a significant difference with the PD's payoff matrix; the Temptation and Sucker's payoff are constantly zero (Osborne, 2004). However, zero payoffs do not reflect how the players should be rewarded in all of the four listed outcomes. A different general sum game could also serve our purpose but it is proven empirically that the PD game is for now the optimum option —was also employed by Banfield (2006), Cleanthous (2010), and Christodoulou et al. (2010).

2.3 Emotions

Emotion is a complex mental construction (Barrett, 2017) that motivates one's thoughts and behavior (Reeve, 2014). Emotions are often indicators of how well the person adapts to threads and challenges of the environment and as a result drive their behavior. However, what causes an emotion? The cognitive arousal theory by Schachter and Singer (1962) which has been highly influential in emotion psychology, describes this process; an emotion occurs when the person feels aroused due to an event that is appraised as concern-relevant and in a particular way, that is specific for this emotion (Reisenzein, 2017). For example, joy occurs when an event (emotion-unspecific arousal) was assessed as a wish fulfillment (emotion-specific cognition). The cognition-arousal theory and its variations are close to the one that we are going now to examine, a theory which is articulated in terms that serve this thesis' purposes.

Computational neuroscientist Edmund T. Rolls adopts the following approach: *"emotions are states elicited by rewards and punishers, that is, by instrumental reinforcers"* (Rolls, 2012). He defines reward as anything for which one will work and, punisher as anything that one will try to escape or avoid. For example, a pleasant emotion might be the happiness produced by reward like a warm hug, or an unpleasant emotion might be the frustration produced by the omission or termination of an expected reward, like the death of a loved one. Moreover, the omission or termination of a punishing stimuli elicits emotions, like the emotion of relief. Note the correlation with the cognitive arousal theory; the concept of appraisal involves assessing a stimuli as rewarding or punishing (Rolls, 2013).

Rolls' above definition suggests that each agent is in an emotional state according to the reinforcement signal it receives. In our case, we might well consider that the values of the payoff matrix, the Temptation (T), Reward (R), Sucker's (S) and Punishment (P) values, are the "rewards and punishers" which cause the agents' emotions. The justification of classifying the emotional states into positive and negative, is given again by Rolls (Rolls, 2012): *"The different emotions can be […] classified according to whether the reinforcer is positive or negative"*. Therefore, given that T and R have positive values, we assume that they elicit positive emotions, whereas S and P have negative values and they elicit negative emotions.

Note that, various events might be the reason that change our current emotional state, and thus influence our self-control behavior. The events might be related or not to the task for which we try to exert self-control. The emotions that are elicited though, due to that event act as internal drivers and thus affect all our decisions, actions, and behavior. Back to our student's dilemma, suppose that an event in his/her life is causing her increasing psychological pain, for instance, a loved one's health is declining, or he/she is victim of bullying. These events are not directly associated with the student's goal of obtaining good grades at the end of the semester, but still impair his/her self-control abilities and might affect his/her overall performance at school. Such an event could be the continuous negative feedback or failure on other school related assignments. The emotions that one

experiences might be caused by unrelated-to-the-task events, but they are internally generated and thus affect the person and his/her behavior in general.

2.3.1 Self-control and emotions

Psychologists suggest that self-control behavior is influenced by one's emotional states. More specifically, one exerts self-control, or any other behavior when his/her emotional state indicates her to do so, since all actions are driven by emotions. The fact that emotion facilitates action has been confirmed by neuroimaging studies which document increased activity in motor areas of the brain during emotional processing of either positive or negative affect (Bremner et al., 1999; Hajcak et al., 2007). In other words, emotions indicate the need that should be satisfied. Pinker (1997) points out that function of emotions: "freely behaving robots […] will have to be programmed with something like emotions merely for them to know at every moment what to do next".

Functional Magnetic Resonance Imaging (fMRI) studies confirm that while anticipating a significant outcome, there is activation in parts of the limbic system which correspond to emotional arousal, providing a framework for understanding how future rewards influence the dilemma (Knutson & Greer, 2008). Not only in the thought of an upcoming reward, there is emotional arousal, but there is also evidence that the experience of intense emotions (positive and negative) interferences and limits the available cognitive resources and thus, the probability that one will make poor decisions inconsistent with her long-term goals increases (Cyders & Smith, 2008).

However, the notion of emotions in the existing models is absent (Banfield, 2006; Cleanthous, 2010; Georgiou, 2015) and their presence would make the model more realistic and will enable us to examine their role in self-control behavior. For this reason, we next examine the existing evidence for the impact which specifically positive and negative emotions, have on self-control.

2.3.2 Self-control and negative emotions

It is easier to first examine the case of negative emotions. When someone is in a bad mood (a negative emotional state) her first concern becomes how to change her mood and as a result to choose the small immediate reward over the large future one. This pattern can continue since from the moment the SS reward is obtained, the individual has again exactly the same options as before the moment she chose (Rachlin, 2000), and there is again the desire to change their mood by choosing the SS reward and not self-regulate. That vicious cycle is the root of addictive behaviors, according to Rachlin (2000). That kind of behaviors, such as alcohol seeking behavior and drug abuse, are more likely to be amplified when the individual is experiencing extreme negative emotional states (VanderVeen et al., 2016). The alcohol related behaviors are in fact associated with impulsive risk-taking and rash decision-making during increased emotional reactivity, conditions which have been identified as main causes of self-control failure (Cyders & Smith, 2008). All in all, extreme emotions lead to extreme actions that are not aligned with the LL reward, but provide immediate reinforcement and thus, are more likely to be repeated in the future (Fischer et al., 2005).

Not only risk-taking behavior is associated with self-control failure though. Studies in children and adults showed that they become more inclined to seek immediate gratification when they feel negative emotions (Robinson et al., 2013). For example, Ruderman (1985) found that women on diet, would eat more crackers when they receive negative feedback that indicates failure, than when they receive successful feedback (Ruderman, 1985). These findings are related to the idea behind hedonic emotion regulation; if you feel bad, you tend to do something to feel better (Robinson et al., 2013). Several more studies showed that negative emotions such as anger, anxiety, fear and sadness often reduce self-control (Cyders & Smith, 2008; Heatherton, 2011; Schmeichel & Tang, 2015). Ultimately, this phenomenon is attributed to the inability of the prefrontal cortex (PFC) to regulate the brain regions (e.g., the amygdala which is part of the limbic system) which promote the negative affect (Heatherton, 2011). More recent studies attempted to explain in more detail the role of the PFC in the case of self-failure due to the presence of negative emotions, and found out that the excessive recruitment of the PFC in the long term appeared to predict self-regulatory failures (Chester et al., 2016;

Knoch & Fehr, 2007). This supports the theory by Baumeister et al. (2004) about self-control, which suggests that self-regulation "operates on the basis of a limited resource".

The effect of negative emotions on self-regulation, is far more complex though. Negative emotions have proven to be necessary and having positive impacts on the self-regulating processes (Robinson et al., 2013). In general, negative emotions do lead to problem-solving action, especially when one faces important problems which needs to address (Hajcak et al., 2007). For this reason, and compounded that humans recognize that they are myopic, they are trained via parenting and schooling to experience immediate negative emotions such as guilt and fear when they succumb to various temptations (Loewenstein & O'Donoghue, 2006). The immediate negative emotions one feels in this case, "replace" the future negative consequences of not focusing on the future reward and facilitates reparative action. The thought that one will feel guilty if she overeats or overspends, is a source of exerting self-control. However, when self-control is not achieved, these negative emotions impose costs with no corresponding benefits (e.g. she continuous to overeats *and* she is gaining weight).

Researchers have studied self-control in social contexts too, where emotions have a very important role (Heatherton, 2011). The negative emotions that one experiences in a social context, like feeling guilty, might help them to exert self-control (R. F. Baumeister et al., 1994). For example, feeling socially excluded, motivates behavior to repair social relationships or feeling ashamed in the thought of cheating our partner helps reign in temptations.

2.3.3 Self-control and positive emotions

Now we turn to examine how positive emotions affect self-control. Baumeister et al. (2007) suggest self-control is a limited resource and on that basis, positive emotions' main effect in the self-regulatory processes appear to be restorative; that is, they recharge the self's resources, enabling people to function effectively again. Empirical studies showed that positive emotion helped to avoid this phenomenon, which was named 'ego depletion', by replenishing the resource and thus enable people to exert self-control (Ren et al., 2010; Tice et al., 2004). This approach is reminiscent of the strength model of self-

control (Roy F. Baumeister et al., 2007), which Cleanthous (2010) used for one of his simulations and achieved to increase the frequency of CC states. The strength model of self-control (Roy F. Baumeister et al., 2007) suggests an analogy between the self-control behavior and a muscle; the more exercise the faster the muscle depletes, but repeated exercise strengthens the muscle in the long-term. Indeed, the effect of positive emotions on self-control and motivation are researched in different domains. For instance, it has been found that positive emotions, in contrast with negative ones, correlate well with language learning motivation (MacIntyre & Vincze, 2017), and successful weight maintenance (Robertson et al., 2017). There is also evidence, that mild increment of positive emotions improves problem-solving skills, such as cognitive flexibility (Isen, 1987), which might be proven useful during a self-control dilemma.

We have mentioned in the section of negative emotions that extreme emotions lead to extreme actions. Extreme emotions though, are not limited to negative ones. In the same way that one might focus on the small sooner (SS) reward when in an extreme negative emotional state, one might also focus on the SS reward when in an extreme positive emotional state (Cyders & Smith, 2008) and take an ill-advised decision. For example, one might get drunk during a celebration or destroy a relationship due to excessive self-confidence by using degrading language. There is also evidence that positive emotions might lead to risk-taking behaviors such as, drug use and gambling (Holub et al., 2005). Therefore, individuals during positive emotional states are likely to fail self-control due to the temporary loss of cognitive control and the false assumption that positive outcomes will result from their actions (Dreisbach, 2006; Nygren et al., 1996). Psychologists in fact, identify the risk-taking behavior which takes place *frequently*, as a sensation-seeking behavior (Cyders et al., 2007).

2.4 Reinforcement Learning

Reinforcement Learning (RL) is a class of algorithms which allow an agent to reach its goal by learning a behavior through the interaction with a dynamic environment and where the only source of feedback is a reward signal (Otterlo & Wiering, n.d.). This concept underlies all human learning theories. Humans do not always learn under supervision. For example, an infant attempts to reach an object or move around by

connecting the consequences of its actions to the results and re-evaluating them in order to achieve its goal. However, not only through childhood but throughout our lives, we constantly learn new skills which are our goals, by just observing the changing environment and figuring out how to act in order to influence the state of the environment and reach our goal. Therefore, RL is a problem faced by an agent who learns how to maximize a numerical reward signal in the long-run while interacting with an environment.

Before we delve more into understanding RL, it is important to note the differences of RL with the other two main sets of techniques in the field of machine learning: the supervised and unsupervised learning. Supervised learning is learning through a labeled set of data. The goal is to construct a predictive model and the desired pairs of actions and outcomes are provided by the labeled data. This is an important mode of learning, but is an impractical way when the environment is dynamic. Unsupervised learning's goal is to construct descriptive models, that is, to reveal patterns and insights in the data. Like in reinforcement learning, the model is not given a target, but it is wrong to assume that the algorithms of both categories achieve the same results. The goal of reinforcement learning is not to unveil a structure, but to maximize a reward signal.

In RL, the agent does not know a priori an optimal policy and is not told which actions to perform. For this reason, the agent has to explore the environment and learn by trial-and-error. At some point in time, the agent will form a policy which then has to try out and evaluate the outcomes. When the actions that are chosen based on that policy turn the agent away from its goal, the agent needs to keep exploring and improve its policy. This is called the exploration-exploitation trade-off; in order to learn it has to explore, but in order to perform well it needs to exploit what it already knows (Woergoetter & Porr, 2008). Balance between exploration and exploitation, is generally difficult, and there are various solutions. We describe how we approach this problem in section 2.4.2.

In order to define the interaction between a learning agent and its environment in terms of states, actions and rewards, the formal framework of Markov decision processes (MDP) is used. Sequential decision-making problems where actions influence subsequent situations are formalized by deploying MDPs. Figurer 2.5 shows an example of the

interaction of an agent with the environment. The agent chooses an action *a* according to the available information about the environment's current state. The action that is taken changes the environment, and the agent perceives that change in the form of the environment's new state and a reinforcement signal. The agents' rewards are the basis for evaluating its choices. Since the agent's objective is to maximize the accumulated reward over time, it forms a policy by which the agent selects actions as a function of states. A *policy* is optimal when the value function assigns to each state-action pair the largest expected return.

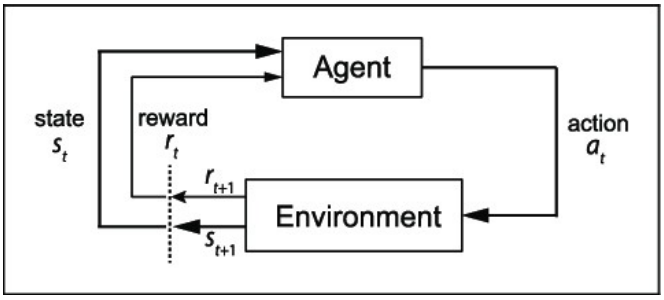

Figure 2.5: *The reinforcement-learning model. At each step t, the agent selects an action a_t according to the state s_t and the reward r_t. The action will then have an impact on the environment from which the next state s_{t+1} and reward r_{t+1} will be induced.*

In our case we deploy *temporal-difference* (TD) learning (Sutton, 1988). TD combines the benefits of both Dynamic Programming (DP) and Monte Carlo methods. TD methods learn directly from raw experience, like Monte Carlo methods, and they update estimates without waiting for the final outcome (they bootstrap) like DP. The update rule of TD is equation (1):

$$V(S_t) \leftarrow V(S_t) + \eta \left[R_{t+1} + \gamma V(S_{t+1}) - V(S_t) \right] \qquad (1)$$

where η is a step-size parameter, γ is the discount factor and R_{t+1} the new reward. We notice that the new estimate is based on the old estimate. Next, we will consider the Q-Learning algorithm which its update rule is a variation on the TD learning rule (1).

2.4.1 The Q-Learning Algorithm

One of the classic model-free algorithms for reinforcing learning from delayed reward is the Q-Learning algorithm (Watkins, 1989). It is one of the most basic methods to estimate Q-value functions and its update rule is the following equation (2):

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \eta \left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t) \right] \qquad (2)$$

where $Q(S_t, A_t)$ is the new Q-value of the state-action pair based on the sum of two parts at the end of each step. At time t+1 it makes a useful update by using the observed reward R_{t+1} by taking action A_t state S_t and the discounted estimate of optimal future reward. That is, the term $\max_a Q(S_{t+1}, a)$ returns the maximum Q-value in the next state S_{t+1} over all actions *a.* This future reward is discounted by the parameter $\gamma \in [0, 1]$ which defines whether the agent is myopic and focus more on the immediate or long-sighted and focus more on the future rewards. Values of the γ parameter closer to 0 indicate that the agent focus on the immediate (or SS) rewards, whereas values closer to 1 indicate that the agent focus on the future (or LL) rewards. The η term is the learning rate, which determines to what extent we weigh the two parts of the sum into the new Q-value.

Taking the maximum across all actions *a*, makes learning independent of the starting policy and allows keeping this policy throughout the whole learning process. This is the reason Q-Learning is an off-policy TD control algorithm, in contrary with the *state, action, reward, state, action* (SARSA) method (Rummery & Niranjan, 1994) which continuously updates its policy during learning (on-policy update). In other words, the Q-Learning update rule guarantees that the optimal policy and the optimal value function are found, a property that makes convergence control much easier.

The effectiveness of the Q-learning algorithm when applied to nonspiking neural networks in order to simulate self-control was first demonstrated by Cleanthous (2010). Later, a simple Q-Learning model as well as the SARSA method were deployed by Georgiou (2015) to simulate the self-control behavior and examine its relationship with consciousness. Georgiou's (2015) results, showed that the agents learnt to exercise self-control more easily with the Q-Learning algorithm in comparison with the SARSA

algorithm. The SARSA method had good results only when it was combined with hill climbing techniques. For the purposes of this thesis we will only deploy the Q-learning algorithm, since it is way more effective, as previous work revealed, and it is more easily handled.

2.4.2 The ε-greedy policy

Along with the Q-Learning algorithm, we need to choose a policy that will allow the agents to decide whether they will explore new moves and payoffs or exploit what they already know about their dynamic environment. This problem is defined as the exploration-exploitation problem, and a good policy that resolves this dilemma, is the ε-greedy policy. This policy tries to balance how the agent chooses which tactic to follow at each round based on the epsilon value (ε). The agent chooses the exploration tactic which might result in a better (still unknown) outcome with probability ε, and the exploitation tactic that uses the already accumulated knowledge with probability $1-\varepsilon$. In other words, the epsilon value $\varepsilon \in (0,1)$ specifies the probability in which the agent is going to explore. For example, if $\varepsilon=0.1$, then the agent will explore new actions with probability 0.1, and it will exploit its knowledge with probability 0.9.

Chapter 3

Design and Implementation

3.1 Introduction

In this chapter we first explain how the presence of emotions are simulated in the self-control model. Positive and negative emotions are simulated according to which combination of payoff values we are changing in the payoff matrix throughout the rounds of the IPD game. Secondly, we present the development of the Q-Learning algorithm which is entirely based on Georgiou's (2015) work, who created two main classes (*Player* and *MainProgram*) using the Object Oriented Programming language Java.

3.2 Simulating emotions

Previous research showed that irrespective of the implementation of the computational model of self-control, that is, whether it is more biologically realistic or not, the results remain the same (Christodoulou et al., 2010). Based on that, we will deploy a Q-learning model in which the two subagents will learn to cooperate in the presence of positive or negative emotion.

We will base our design of the simulation on Rolls' (2012) definition on emotions, that was presented in section 2.2. Since reinforcement signals represent the presence of emotion, we will gradually change the values of the payoff matrix, in order to magnify the effect of positive (T, R) and negative (P, S) reinforcing signals. The initial hypothesis is that positive emotions promote self-control, whereas negative emotions impair it (Tice et al., 2004). Therefore, the changes on the reinforcing signals should reflect in some experiments, the increase of positive emotions, and in others the increase of negative emotions.

Operant conditioning (or firstly referred to as instrumental learning) is the theory behind techniques that psychologists use in order to change (non)-human behavior and which inspired us on how to simulate the presence of positive and negative emotions. In particular, operant conditioning is an associative learning technique based on rewarding or punishing (omission of reward) a certain behavior by adding or removing stimuli (Skinner, 1938). We have defined the values of the payoff matrix (T, R, P, S) as the reinforcing signals that elicit emotional states. In the same way, we could also consider them as stimuli which are added or removed with the goal to enhance or weaken a behavior. The combination of the goal and the method used produces the four components of operant conditioning as they are shown in Figure 3.1. Positive and negative reinforcement have the goal of increasing self-control, and thus we consider them as ways one simulates the presence of positive emotions, whereas positive and negative punishment's goal is to reduce self-control. The simulation of the presence of positive emotions is done in the two following ways:

a. The increment of the R payoff (positive reinforcement).
b. The decrement of the T payoff and the increment of the P and S payoffs model the elimination of negative emotions (negative reinforcement). Let me remind here, that the P and S payoffs are negative values (< 0) and therefore, increment. Now, the simulation

The simulation of the presence of negative emotions is done in the two following ways:

a. The increment of the T payoff and the decrement of the P and S payoffs models the increment of the presence of negative emotions (positive punishment).
b. The decrement of the R payoff models the elimination of positive emotions (negative punishment).

	Add stimuli	Remove stimuli
Increase Behavior	Positive Reinforcement	Negative Reinforcement
Decrease Behavior	Positive Punishment	Negative Punishment

Figure 3.1: *Simulating the presence of emotions based on the components of operant conditioning: positive reinforcement, negative reinforcement, positive punishment and negative punishment.*

According to Rolls (2012), *"we are sensitive to some extend not just to the absolute level of reinforcers being received, but also to the change in rate and probability."* Therefore, we could experiment with the *ratio* in which we will increase or decrease the values of the payoff matrix, over the rounds, as well as with the number of *rounds* that will elapse before changing the values again. The significance of these two parameters will be demonstrated in the conclusions. Next, we will further elaborate on how these changes can take place while always having in mind that the two rules of the IPD game (section 2.3) must never be violated.

3.2.1 Positive emotions

The presence of positive emotions can be reflected in the game by increasing the Reward value for mutual cooperation and keeping the rest of the values the same. An increased Reward would motivate the agents to reach the CC state faster. In other words, a positive emotion motivates reaching the LL reward, according to the way self-control was defined earlier (positive emotion corresponds to LL and not the SS reward). The rules of the game must always be satisfied, though. This means that the Reward must never exceed the Temptation value (1st rule), otherwise it would not be a PD game anymore. Thus, we expect to see the overall accumulated payoff to increase and approximate the theoretical best, while the frequency in which we reach CC states, to be greater.

Increasing the Reward value can be considered as a form of positive reinforcement, since we add value in order to enhance a certain behavior. In the same pattern, we wondered what would happen if a kind of negative reinforcement was applied. For example, decreasing the Temptation value, can be a considered as negative reinforcement, since we remove some value in order to increase the self-control behavior. A decreased T value that is getting closer to the Reward value, would make the option to defect less attractive and thus, cause an increase in the CC states. Nevertheless, we do not anticipate that this method will surpass the first one, but rather having significantly less impact.

It should be noted that the last-mentioned method does not increase the positive reinforcing signals and thus adding positive emotion, but rather *decreases* the positive reinforcers of *each* of the agents. This statement seems to contradict what we have defined as the "presence of positive emotion" (increasing the positive signals). However, we need to also examine the effects on self-control, when we use methods that simulate the elimination of the negative emotions. Note that the decrement of the Temptation payoff simulates the decrement of positive emotions for *each* of the agents, and the decrement of negative emotions for *both* agents. Decreasing the Temptation, eases the internal conflict that causes negative emotions (Schacht & Sommer, 2012). Cleanthous (2010) was the first that indicated that the increment of the Temptation (T) payoff, increases the internal conflict and the decrement of T, decreases it by deploying payoff matrices of different internal conflict intensities (e.g., "moderate", "strong"). After all, self-control is proven to be enhanced not only by the presence of positive emotions, but also by the elimination of negative emotions. According to the hedonic psychology theory, the assessment of subjective well-being, consists of three components; life satisfaction, the presence of positive mood, and the absence of negative mood, "together often summarized as happiness" (Ryan & Deci, 2001). Therefore, the *absence* of negative mood should also be considered as a way of nourishing positive emotional states.

In the same way that decreasing the Temptation payoff eliminates negative emotions and contributes to a more positive emotional state, we suggest the following methods which are concerned with the negative reinforcers. The goal now, is to explicitly restrict the effect of the negative payoff values (Punishment and Sucker's). In order to do that, we have to add a positive ratio to the P and S values.

The combination of the methods that contribute to the absence of negative emotions is also tested. That is, except from testing each one of the last three suggested methods separately (decreasing T, increasing P, increasing S), the combination of them is a good way to "boost" the effects of these changes on the final outcomes. We have tested the combination of increasing both the P and S values, as well as the combination of all three of them.

3.2.2 Negative emotions

The negative emotional states can be modeled in several ways. According to our definition, the negative reinforcers elicit the negative emotions, and therefore, it is expected that the agents will fail in achieving self-control behavior. The negative reinforcers are the Punishment (P) and Sucker's (S) payoffs. The effects of increasing their absolute value (or decreasing them since we add a negative ratio), depend on whether those methods is used in combination with others. In the case of only decreasing the Sucker's payoff value, we expect that the model will not converge in the CC state, but in the CD or DC states, since one of the agents will constantly receive the highest reward which is the Temptation payoff. Although we expect to experience impairment of self-control by only decreasing S, it is not the case with only decreasing the Punishment payoff. Regardless of how much we decrease P, we expect that the agents will learn that the best outcome is the CC state. There is also the option to decrease the P and S values at the same time.

As we mentioned earlier in section 3.2.1, the increment of the Temptation value, also increases the level of conflict, and thus, the negative emotions that the system experiences. For this, increasing the Temptation payoff is another method that is tested. Adding a negative ratio to negative reinforcers (decreasing P or S) and adding a positive value to a positive reinforcers in order to decrease a behavior is a form of positive punishment, in analogy to the positive reinforcement that will be used to increase self-control (increasing R). Therefore, the combination of increasing the Temptation value and decreasing the Sucker's payoff value simultaneously, would make it even more tempting to defect and the impairment of self-control even more obvious. Both rules of

the IPD game should always be satisfied. Despite that the 1st rule allows us to increase the Temptation payoff as much as we want, the 2nd rule demands that the sum of the T and S values do not exceed two times the R value. This makes sure that the dilemma holds all the time; if the T value could increase infinitely, one would know that this is the best outcome for them and there would be no need of exercising self-control.

Another approach for modeling negative emotions is to decrease the Reward payoff, and thus making it less tempting to cooperate. The Reward and Temptation values' difference will increase, so we expect that the system will at least experience a drop in the frequency of the CC states. This approach can be thought as a negative punishment approach since we "remove" the reward in order to decrease the behavior of self-control. It will be used in combination with decreasing the S and/or P payoff values.

3.3 The Q-Learning agents

The *Player* class is used to create a Q-Learning agent. In our case, we have two agents who represent the upper and lower parts of the brain and who are playing the IPD game. Each agent has its own parameters; the learning rate (η), the epsilon value which defines the ε-greedy policy, the discount factor (γ), the Q-table which holds the Q-values that are estimated for every state-action pair, the payoff matrix, the action that the agent chooses at each step. The agents are objects created at the beginning of the program and are of type *Player*. In the Player's constructor, the Q-table and the payoff matrix are initialized. There are 4 states which the agents can result in (CC, CD, DC, DD) and 2 actions from which they can choose, they can either cooperate (C) or defect (D). Thus, the Q-table's dimensions are 4 rows by 2 columns. Parts of the class *Player* are also methods for setting and getting the values of the object's parameters.

The ε-greedy policy that was chosen as solution to the exploration-exploitation problem (section 2.4.2) is implemented in the method *choose_action.* Finally, the class *Player* has the method which constitutes the heart of the Q-Learning algorithm; the *update_Q* method which implements the update rule of Q-Learning (section 2.4.1). This is the method responsible for updating the Q-values of the Q-table from which the agent learns throughout the rounds.

3.4 The Q-Learning model

The Q-Learning model is built in the class *MainProgram*. Every time the *MainProgram* is executed, the model runs for 35 trials of 1000 episodes each. First, we initialize 3 arrays which will hold during the agents' learning the states that the agents reach, the rewards each agent earns from the payoff matrix, as well as the overall payoff of the agents. When the learning is done, the data of that arrays will be processed and saved in text files. Each agent is initialized through the constructor of the object *Player* (section 3.3). The values T, R, P, S of the payoff matrix of both agents are set, as well as the epsilon value of the ε-greedy exploration. The discount factor (γ) is set to 0.1 for the agent that represents the lower part of the brain (the limbic system), which makes the agent focus on the SS rewards, whereas for the agent that represents the higher part (the prefrontal cortex), is set to 0.9 which makes the agent focus on the LL rewards.

Before the agents start learning, the five parameters that define how the values of the payoff matrix are going to change throughout the game are set; the ratios for the Punishment value (*ratio_P*), the Reward value (*ratio_R*), the Punishment value (*ratio_P*), the Sucker's value (*ratio_S*), and the number of rounds (*rounds*) between of each change which is the same for all ratios. The ratios can be positive or negative values. The update of each value of the payoff matrix takes place if and only if the two rules (1) and (2) of the IPD game are satisfied (section 2.3), and after the number of rounds that we specified have elapsed.

At each round, both agents select an action given their current state, which is used to update the Q-table of each agent. During the learning and at each round we save the state which is reached and their returns. After the 1000 episodes, the *payoffs.txt* and *states.txt* files contain the results. In the file *payoffs.txt* each of the three columns is the accumulated payoff of the lower part of the brain, the higher part, and their sum, respectively. In the file *states.txt* each of the 4 columns are the average percentage of the frequency of the CC, CD, DC, and DD states, respectively. Based on these two files we produce the figures that describe the results of Chapter 4. There are four types of figures which depict the effects of the changing payoff values; two of them are about the overall average state

outcomes (CC, CD, DC, DD) and the other two concern the overall performance of the agents, which is measured based on their accumulated payoff:

a. A line plot with 4 traces, one for each of the outcomes from which we can observe the effects of the changing payoff values throughout the rounds. For example, whether an outcome is surpassed over the other and on which round, or how fast the system converges to a certain outcome.
b. A bar plot which shows the overall average outcomes, that is, the percentage of the appearance of each state. The information that we retrieve from this chart is only the difference between the outcomes and which state dominated at the end.
c. A line plot which shows the overall accumulated payoff (this is the 3rd column in the *payoffs.txt*) during the game. The joint performance of the agents is compared with the theoretical best performance (see section 4.1.1).
d. A line plot with two traces, one for each of the agents which are their accumulated payoffs throughout the rounds during the game (1st and 2nd columns in the *payoffs.txt*). Again, their performance is compared with the theoretical best performance.

Chapter 4

Results and Discussion

4.1 Introduction

In this chapter we are going to examine the results of the methods that simulate the presence of positive and negative emotional states, as those were explained in Chapter 3 (section 3.2), while the Q-learning agents engage in the IPD game.

4.1.1 Constant payoff matrix

The results of Figures 4.1a-c are considered the baseline results for our next experiments. The basic model of self-control as was described in the thesis of Georgiou (Georgiou, 2015) sets the learning rate is set to 0.1 and the epsilon value to 0.1. The initial payoff matrix that is used in this and all the experiments that follow, has the values T=5, R=4, P=-2, S=-3. The values of the initial payoff matrix correspond to "an internal conflict of moderate intensity" (Cleanthous, 2010). The program runs 30 trials of the game and each time the IPD game has a duration of 1000 rounds. There are four types of figures; Figure 4.1a shows the average of the outcomes during the IPD game, Figure 4.1b depicts the average of the outcomes after the 1000 rounds and Figure 4.1c shows the overall performance of the Q-learning agents, as well as the performance of each of the agents.

Accumulated payoff we define as the sum of the values of all the reinforcing signals, that is, the values of the payoff matrix, that an agent receives during the 1000 rounds. The overall performance (Figure 4.1c *left*) derives from the addition of the accumulated payoffs of each agent at each round, and then we find the average accumulated payoff when we divide with the number of the trials. The theoretically best performance is shown for comparison (dot-dashed line). The theoretical maximum that corresponds to playing CC all the time and thus receiving R + R = 4 + 4 = 8 for each round is 8000 (8 × 1000). The performance of each agent (Figure 4.1c *right*) derives from adding the reinforcers that it gets at each round, and then calculating its average accumulated payoff according to the number of trials. The theoretically best performance is shown for comparison (dot-dashed line). The theoretical maximum that corresponds for an agent to defect all the time and thus receiving T = 5 for each round is 5000 (5 × 1000).

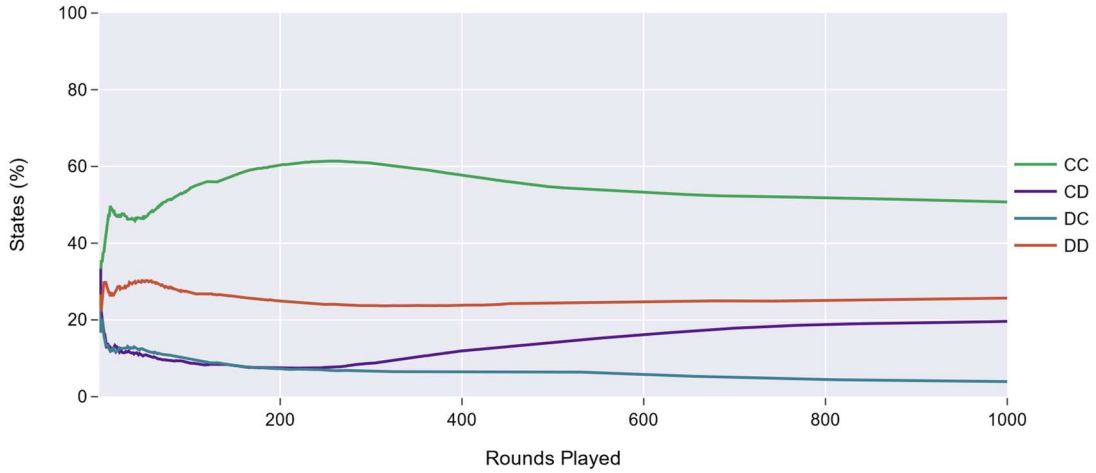

Figure 4.1a: *Average of the outcomes CC, CD, DC and DD during 1000 rounds of the Q-learning agents playing the IPD game. The constant payoff matrix is used: T=5, R=4, P=-2, S=-3.*

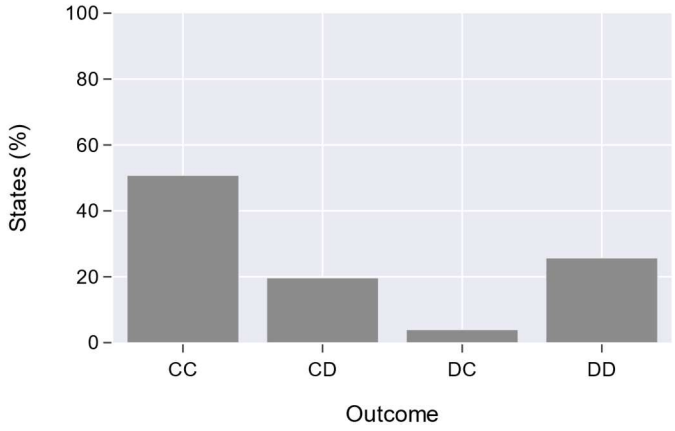

Figure 4.1b: *Overall average outcomes after 1000 rounds of the Q-learning agents playing the IPD game. The constant payoff matrix is used: T=5, R=4, P=-2, S=-3.*

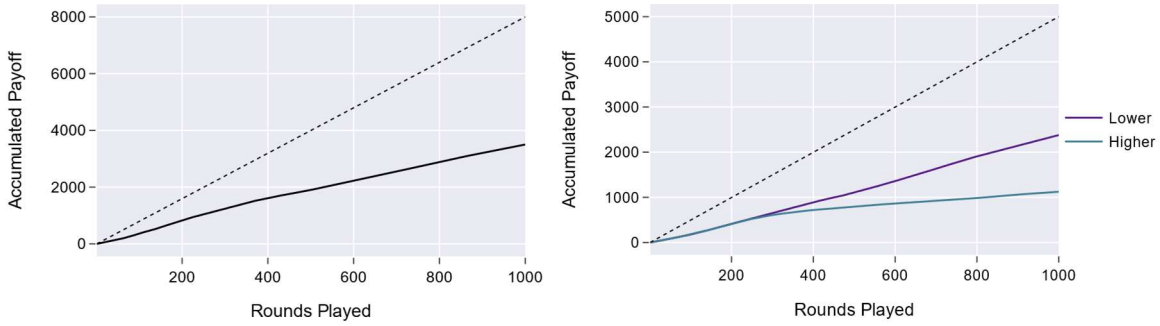

Figure 4.1c: (left) *Overall performance of the Q-learning agents during the 1000 rounds of the IPD game (thick line). The theoretically best performance is shown for comparison (dot-dashed line).* ***(right)*** *Performance of each Q-learning agent during the IPD game. The theoretically best performance for each agent is shown for comparison (dot-dashed line). The constant payoff matrix is used: T=5, R=4, P=-2, S=-3.*

4.2 Simulating positive emotions

Having used a constant payoff matrix in 4.1.1 that produced the baseline results, we are now ready to provide the model with a non-constant matrix, so that we can test our initial hypothesis that positive emotional states improve self-control. In order to simulate the presence of positive emotions, some of the values of the payoff matrix (T, R, P, S) will gradually change during the 1000 rounds of the IPD game. The ways that are tested are increasing the R payoff, increasing the P and S payoff separately as well as at the same time, decreasing the T payoff, and changing the T, P and S at the same time.

The two factors that influence our results while we change the T, R, P, S values are the ratio in which the value changes and the interval (number of rounds) between each change. After experimenting with different combinations of ratio and interval values, we next present our most important findings. We use the same initial payoff matrix, learning rates and epsilon value as in section 4.1.1.

4.2.1 Increasing the Reward payoff

At the first attempt we increase the Reward (R) payoff by 0.1 every 50 rounds. We began experimenting with small ratios like 0.1 because the R value cannot increase more than 1 point since it would exceed the T value and break the first rule of the PD game. However, we were expecting that this small change will increase the total number of CC states. Indeed, the CC states reach the 60% of all the outcomes, around 10 percentage points more than our baseline results (Figures 4.2a-b). We also notice that the overall performance is increased significantly, from 4000 to 6000 and thus, approaching the theoretical maximum accumulated payoff (Figure 4.2c).

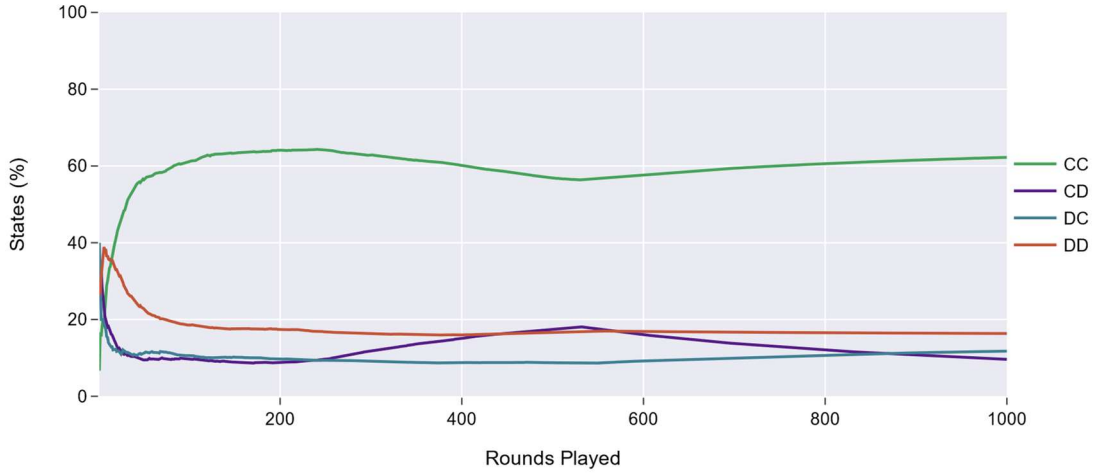

Figure 4.2a: *Average of the outcomes CC, CD, DC and DD during 1000 rounds of the Q-learning agents playing the IPD game. Increasing the R payoff. Ratio=0.1, Interval=50 rounds.*

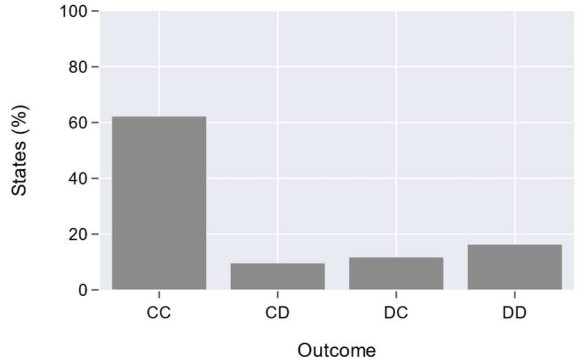

Figure 4.2b: *Overall average outcomes after 1000 rounds of the Q-learning agents playing the IPD game. Increasing the R payoff. Ratio=0.1, Interval=50 rounds.*

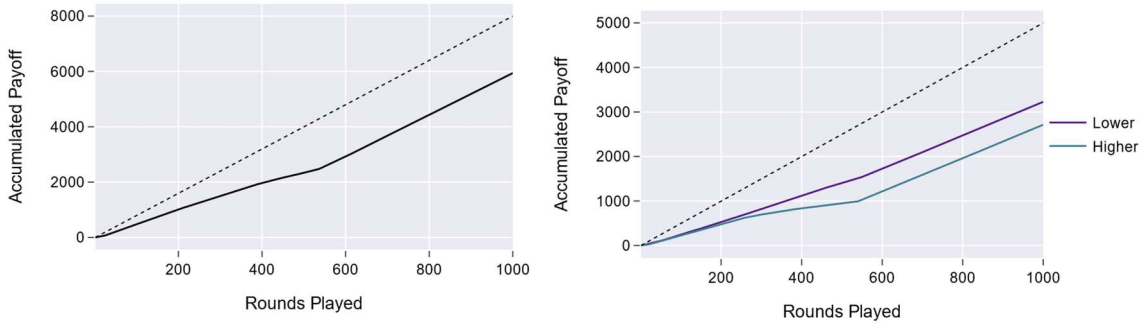

Figure 4.2c: (left) *Overall performance of the Q-learning agents during the 1000 rounds of the IPD game (thick line). The theoretically best performance is shown for comparison (dot-dashed line).* ***(right)*** *Performance of each agent. The theoretically best performance for each agent is shown for comparison (dot-dashed line). Increasing the R payoff. Ratio=0.1, Interval=50 rounds.*

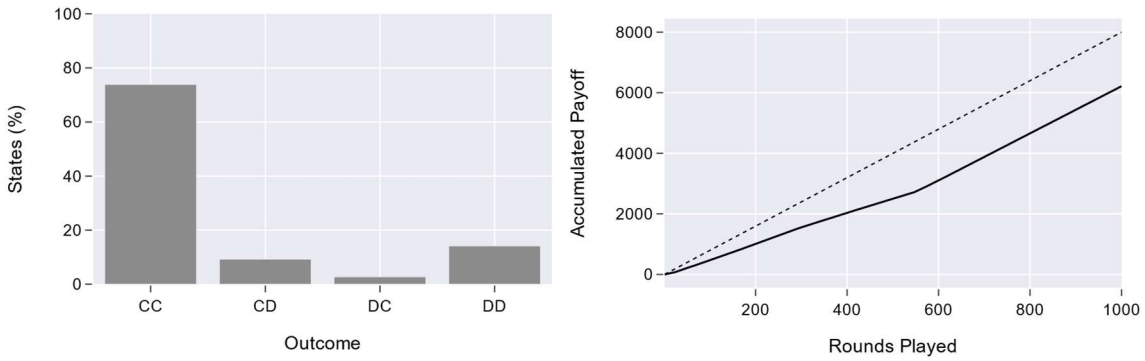

Figure 4.3: (left) *Overall average outcomes after 1000 rounds of the Q-learning agents playing the IPD game. Overall performance of the Q-learning agents during the 1000 rounds of the IPD game (thick line).* ***(right)*** *The theoretically best performance is shown for comparison (dot-dashed line). Increasing the R payoff. Ratio=0.25, Interval=25 rounds.*

In the second attempt we increased the ratio in which the R was increasing to 0.25 and we decreased the interval to 25 rounds; we notice in Figure 4.3 a slight improvement in the number of CC states of 15 percentage points and thus in the accumulated payoff. Therefore, a larger positive change that is given more often (smaller interval) improves self-control behavior.

4.2.2 Increasing the Punishment and the Sucker's payoff separately

As we mentioned before, increasing the P and S values simulates a decrement in the negative emotional states. The ratios that were mainly tested were 0.1, 0.5 for the S value and 0.1, 0.5, 1 and 2 for the P value. The intervals used were mainly 10 and 50 rounds. In contrast with the R and S values, the ratio of P can get values like 1 and 2 since the value difference with the R value that should not reach and thus break the game's first rule, is large enough (4+2=6).

We expect that this kind of simulating the positive affect will not be as effective as increasing the R value, since it is a kind of negative reinforcement. Among all the combinations of ratios and intervals for increasing P or S, the highest overall percentage of CC states (68.2%) which does not exceed the 75% of Figure 4.3, was produced when we increased the P value by 0.5 and every 50. Although the 0.5 ratio did not produce a significantly better result when given more often (interval of 10 rounds), it made a subtle

difference when the interval increased to 50 rounds. This indicates that the rarer increment gives a better result, in contrast with increasing the R where the more frequent the change, the better.

That lead testing a larger ratio using the 50 rounds interval. It was interesting to notice that the larger ratio (1, 2) did not increased the total CC states which reached a maximum of 63.2%, even when it was tested with an interval of 100 rounds. We got similar results as Figure 4.4b shows, when we increased the S value with ratio 0.1 or 0.5 and interval 10 rounds.

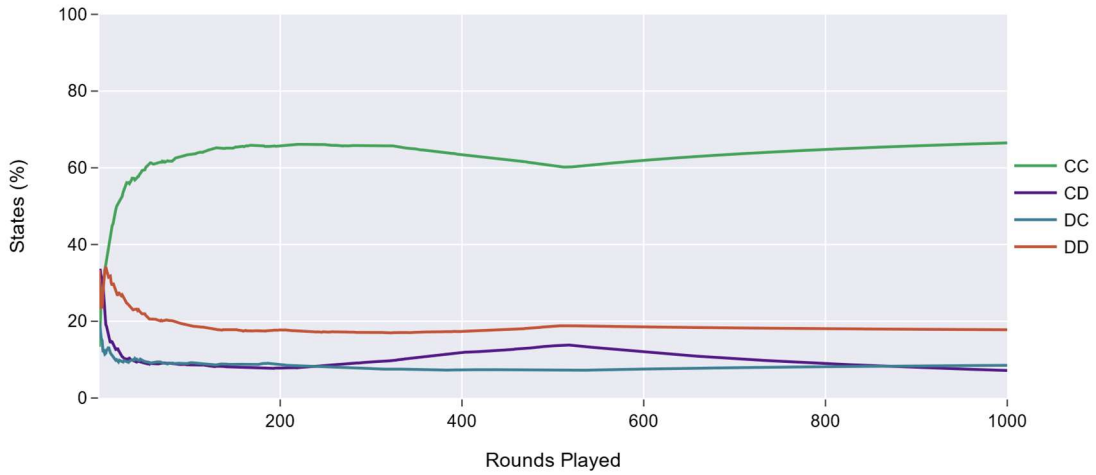

Figure 4.4a: *Average of the outcomes CC, CD, DC and DD during 1000 rounds of the Q-learning agents playing the IPD game. Increasing the S payoff. Ratio=0.1, Interval=10 rounds.*

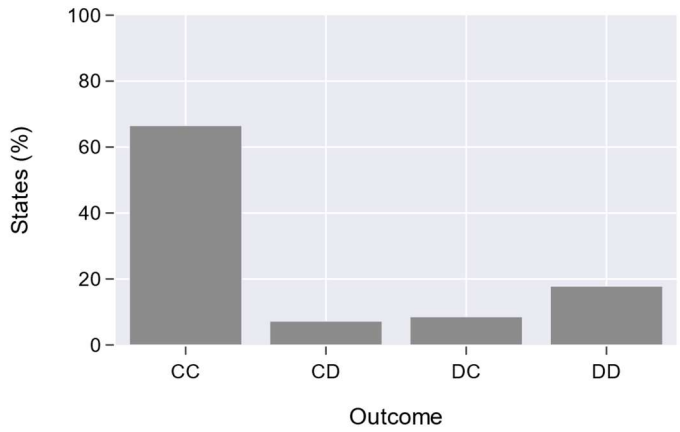

Figure 4.4b: *Overall average outcomes after 1000 rounds of the Q-learning agents playing the IPD game. Increasing the S payoff. Ratio=0.1, Interval=10 rounds.*

4.2.3 Increasing the Punishment and the Sucker's payoff at the same time

We continue testing the hypothesis about the effects of positive affect on self-control by increasing the P and the S payoffs at the same time and expecting that the combination of the two methods will increase the levels of CC states in comparison with the results we produced in the section 4.2.2. However, we notice something remarkable when the ratio of both values was set to 0.1 and the interval to 10 rounds; the self-control behavior is not achieved. As we can see in Figure 4.8, the DD state prevails just until the 600th round, while for the rest of the IPD game, the DC state dominates. Figure 4.5b also shows us that the DC state reached overall 44.8%, followed by the CC state (31.9%) and then the DD state (19.7%).

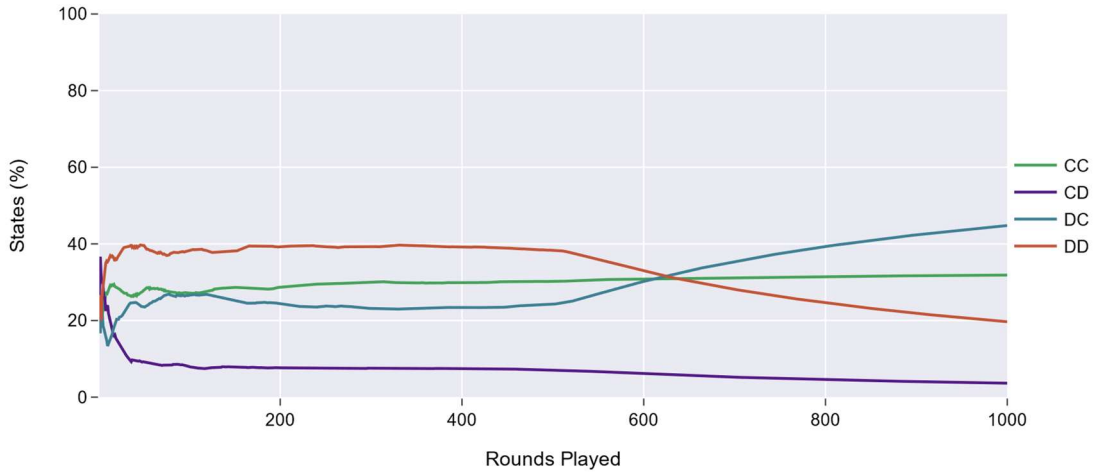

Figure 4.5a: *Average of the outcomes CC, CD, DC and DD during 1000 rounds of the Q-learning agents playing the IPD game. Increasing the P and S payoffs at the same time. Ratio=0.1 (for both), Interval=10.*

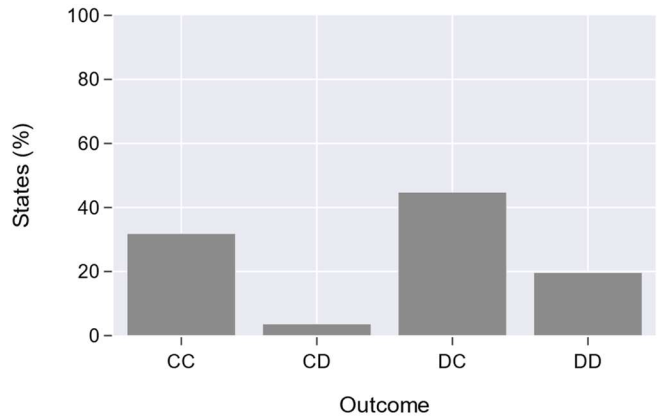

Figure 4.5b: *Overall average outcomes after 1000 rounds of the Q-learning agents playing the IPD game. Increasing the P and S payoffs at the same time. Ratio=0.1 (for both), Interval=10.*

After getting the above result (Figures 4.5a, 4.5b), the expectations changed, and we believed that this combination of methods (increasing P *and* S) would continue to give similar results. However, when the interval changed to 50 rounds, while maintaining the ratio to 0.1 for both values, the agents engaged in self-control (62.3% of CC states) as we see in the Figures 4.6a and 4.6b. The self-control behavior was achieved even when we decreased the interval to 5 rounds and kept the ratio of 0.1. We also continued to achieve self-control, with overall percentage of CC states (61.5%), when we changed the ratio to 0.5 for both values but kept the 10 rounds interval.

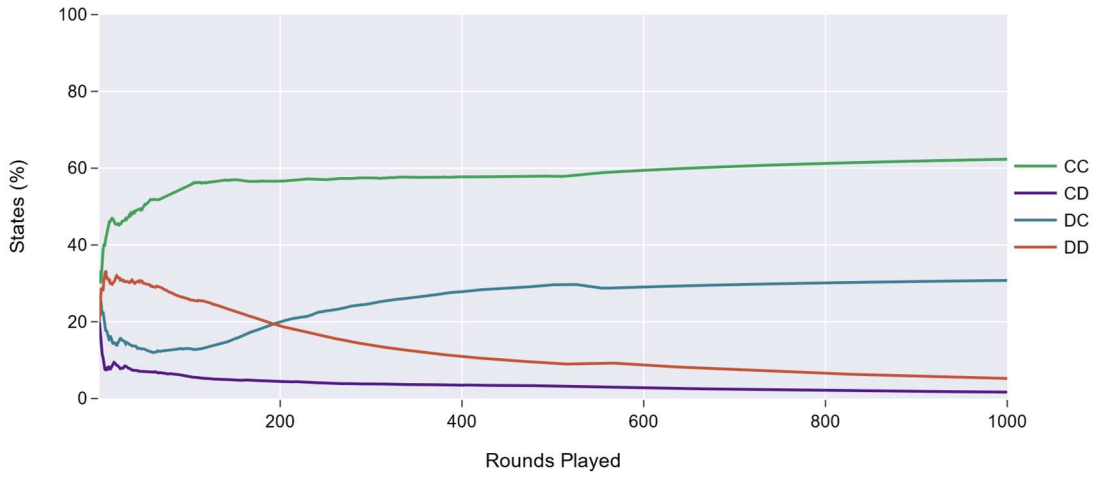

Figure 4.6a: *Average of the outcomes CC, CD, DC and DD during 1000 rounds of the Q-learning agents playing the IPD game. Increasing the P and S payoffs at the same time. Ratio=0.1 (for both), Interval=50.*

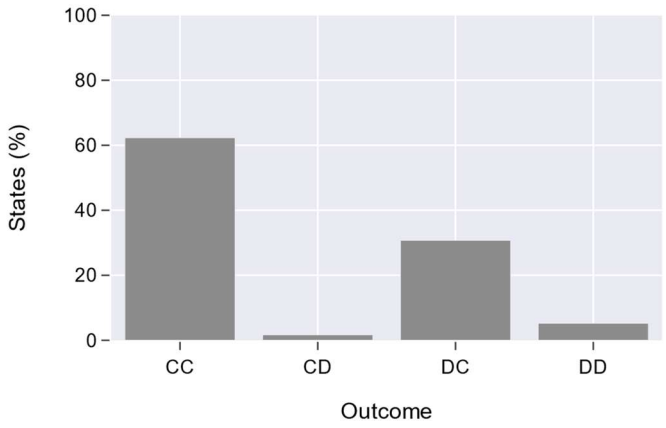

Figure 4.6b: *Overall average outcomes after 1000 rounds of the Q-learning agents playing the IPD game. Increasing the P and S payoffs at the same time. Ratio=0.1 (for both), Interval=50.*

It is noteworthy that no other configuration except the one described above with ratio for P and S at 0.1 and 10 rounds interval, fails to reach the self-control behavior. For example, with ratio for both values at 0.5 and, 10 *or* 50 rounds interval, self-control is still achieved. Moreover, having different ratios for P and S also continued to give positive results. For instance, when the ratio of P was 1 and the ratio of S was 0.1, with 10 rounds interval.

4.2.4 Decreasing the Temptation payoff

Another method that is expected to improve the self-control behavior is the decrement of the T payoff. After testing several combinations of ratios and intervals, the best performance was achieved with ratio 0.1 and 10 rounds interval (Figures 4.7a-c). The overall CC states reached 70.6% and the overall accumulated payoff just under 6000. These results are the second higher after the increment of R by 0.25 every 25 rounds. That indicates small and frequent changes are enough to boost self-control. After all, it is not effective to decrease T by larger ratios since it has only 1-point difference with the R value, which we do not want to reach and break the first rule of the IPD. Despite that the 0.5 ratio with 10 rounds interval was tested and gave 63.3% of CC states. Moreover, we kept the ratio at 0.1 and we increased the interval to 50 rounds, which produced 65.6% of CC states. The two last mentioned attempts are shown in comparison in Figure 4.8.

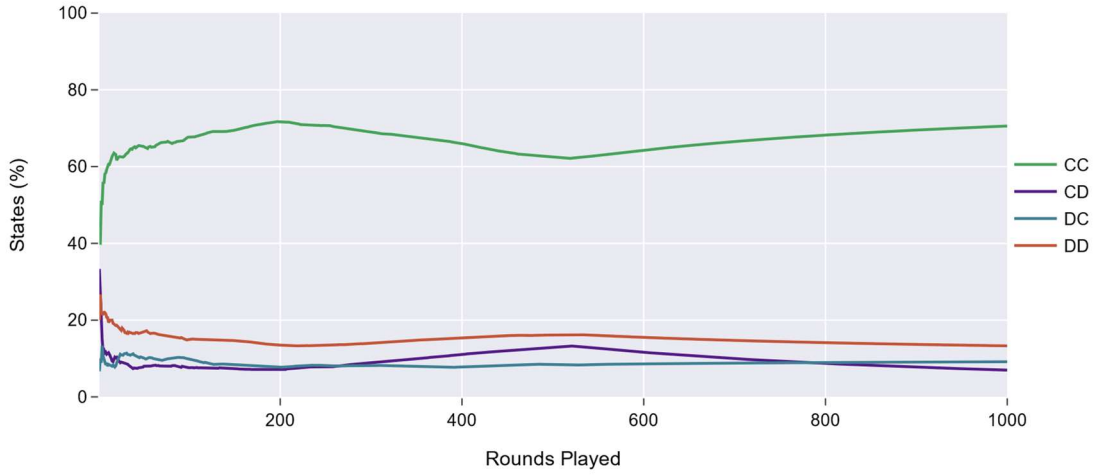

Figure 4.7a: *Average of the outcomes CC, CD, DC and DD during 1000 rounds of the Q-learning agents playing the IPD game. Decreasing the T payoff. Ratio=0.1, Interval=10.*

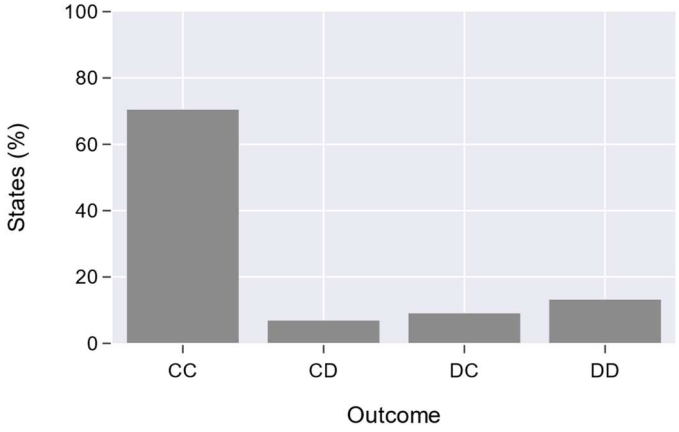

Figure 4.7b: *Overall average outcomes after 1000 rounds of the Q-learning agents playing the IPD game. Decreasing the T payoff. Ratio=0.1, Interval=10.*

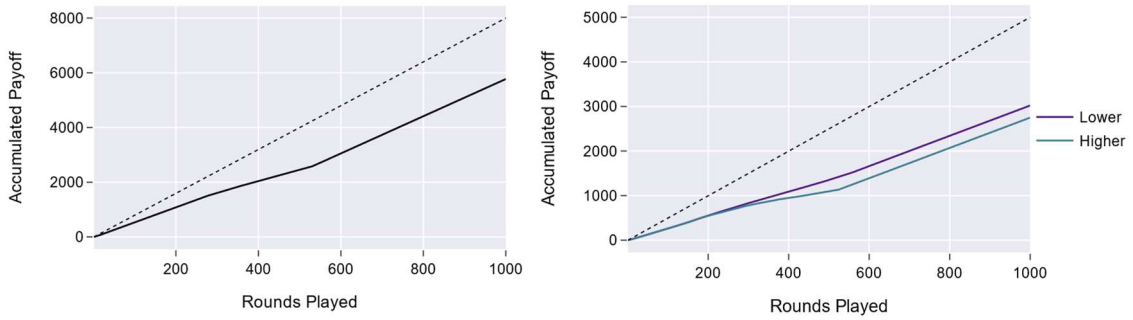

Figure 4.7c: (left) *Overall performance of the Q-learning agents during the 1000 rounds of the IPD game (thick line). The theoretically best performance is shown for comparison (dot-dashed line).* ***(right)*** *Performance of each Q-learning agent during the IPD game. The theoretically best performance for each agent is shown for comparison (dot-dashed line). Decreasing the T payoff. Ratio=0.1, Interval=10.*

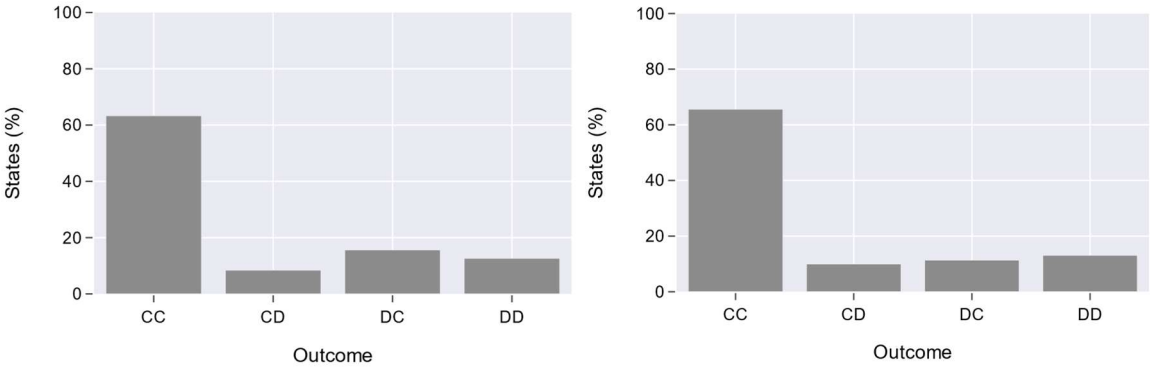

Figure 4.8: *Overall average outcomes after 1000 rounds of the Q-learning agents playing the IPD game.* ***(left)*** *Decreasing the T payoff. Ratio=0.5, Interval=10.* ***(right)*** *Decreasing the T payoff. Ratio=0.1, Interval=50.*

As it was mentioned above, the "problem" that raised using the payoff matrix with values T=5 and R=4 was that we could not decrease the T value by a large ratio due to the small difference with the R value, so we can see the effect of the ratio's magnitude. One might think that decreasing T by using a larger ratio would be more effective, in analogy to the larger ratio that was used to increase the R value in section 4.2.1. For this reason, we set the initial value of T in the payoff matrix to 9 and we made two attempts: using ratio 1 and 50 rounds interval (Figure 4.9) and using ratio 0.1 and 10 rounds interval. Figures 4.10a and 4.10b show the results in comparison.

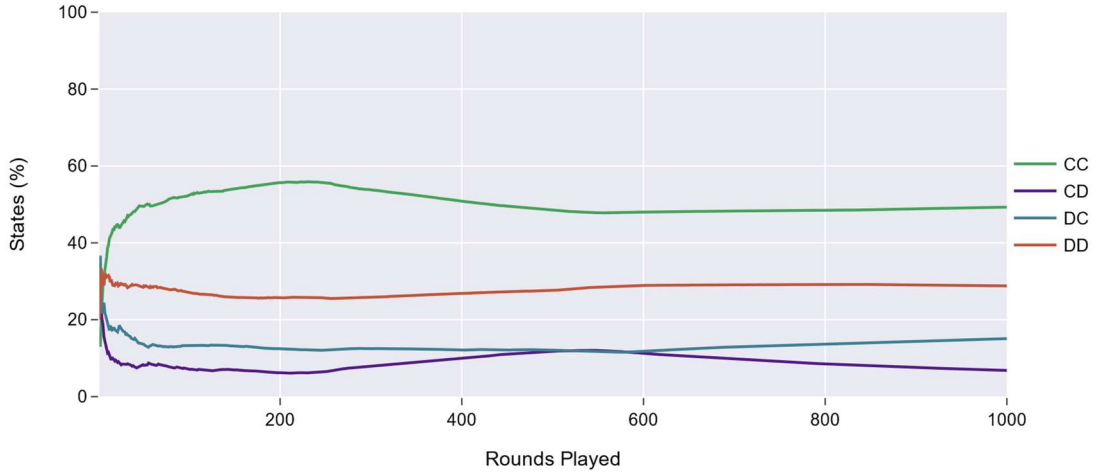

Figure 4.9: *Average of the outcomes CC, CD, DC and DD during 1000 rounds of the Q-learning agents playing the IPD game. Initial payoff matrix T=9, R=4, P=-2, S=-3. Decreasing the T payoff. Ratio=1, Interval=50.*

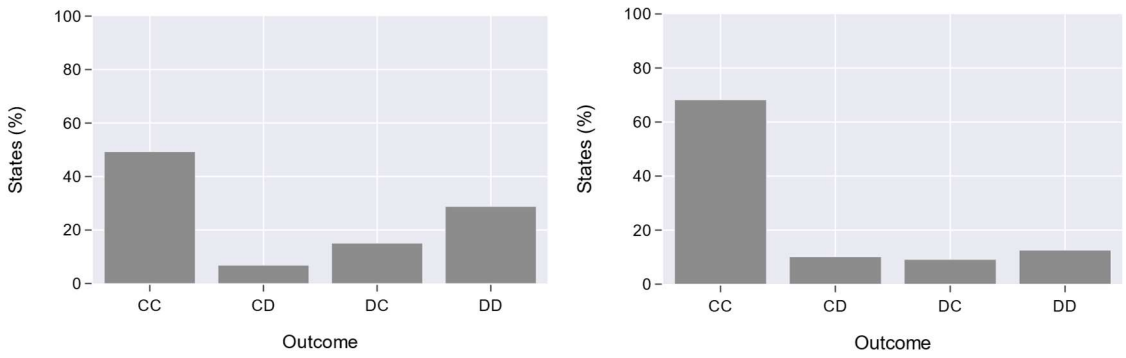

Figure 4.10a: *Overall average outcomes after 1000 rounds of the Q-learning agents playing the IPD game. Initial payoff matrix T=9, R=4, P=-2, S=-3.* ***(left)*** *Decreasing the T payoff. Ratio=1, Interval=50.* ***(right)*** *Decreasing the T payoff. Ratio=0.1, Interval=10.*

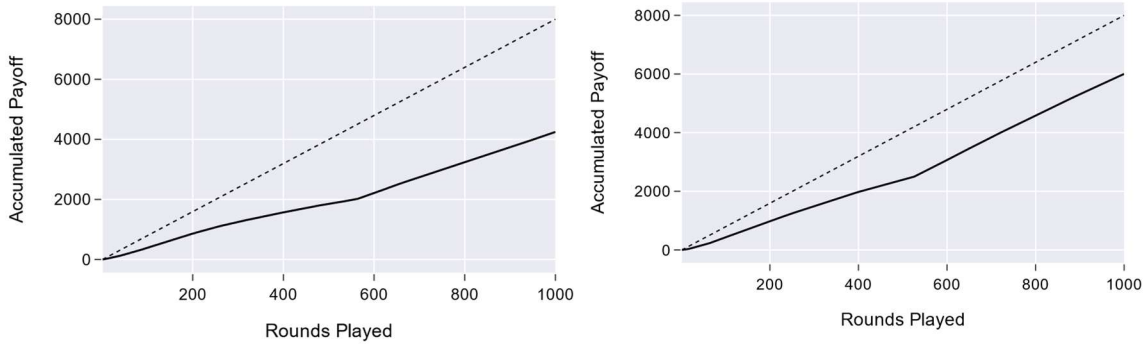

Figure 4.10b: *Overall performance of the Q-learning agents during the 1000 rounds of the IPD game (thick line). The theoretically best performance is shown for comparison (dot-dashed line). Initial payoff matrix T=9, R=4, P=-2, S=-3.* ***(left)*** *Decreasing the T payoff. Ratio=1, Interval=50.* ***(right)*** *Decreasing the T payoff. Ratio=0.1, Interval=10.*

After setting the T value to 9, the results show that the method of decreasing the T value is not only more effective when it is done gradually and in smaller intervals, but also that it is damaging to decrease it using large ratios. Despite that the combination of ratio 1 and 50 rounds interval still achieved self-control, the overall CC states were only the 49.3% (Figure 4.10a *left*) and the overall accumulated payoff was significantly less than our baseline's.

4.2.5 Decreasing the Temptation and increasing the Punishment and Sucker's payoff

The last method that was used to test the effects of the removal of the negative emotions was a combination of the above methods, that is decreasing the T and increasing the P and S, all at the same time. We expect that since we will combine the three methods and will reduce the conflict that the agents experience, to see a rise of the overall CC states. We set the ratio to 0.1 for all the three values, since for T and S was the ratio that gave the best performance in the experiments above and for P is a reasonable ratio that will not break the game's first rule. Moreover, 10 and 50 rounds interval were used since they would not affect the performance. Figures 4.11a-c show the best performance of all methods regarding positive emotions that was achieved using this method.

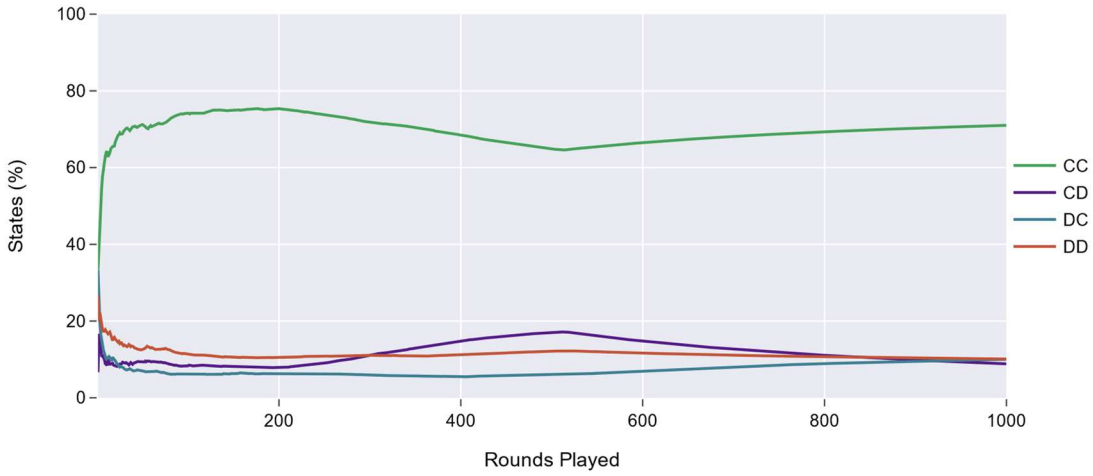

Figure 4.11a: *Average of the outcomes CC, CD, DC and DD during 1000 rounds of the Q-learning agents playing the IPD game. Decreasing the T payoff and increasing the P and S payoffs. Ratio=0.1 (for all the payoffs), Interval=10.*

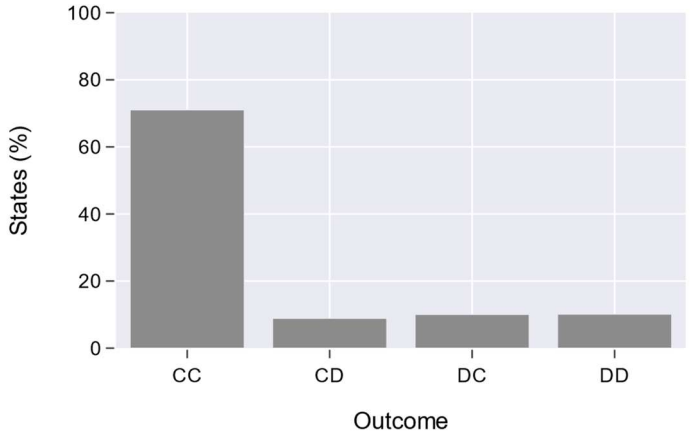

Figure 4.11b: *Overall average outcomes after 1000 rounds of the Q-learning agents playing the IPD game. Decreasing the T payoff and increasing the P and S payoffs. Ratio=0.1 (for all the payoffs), Interval=10.*

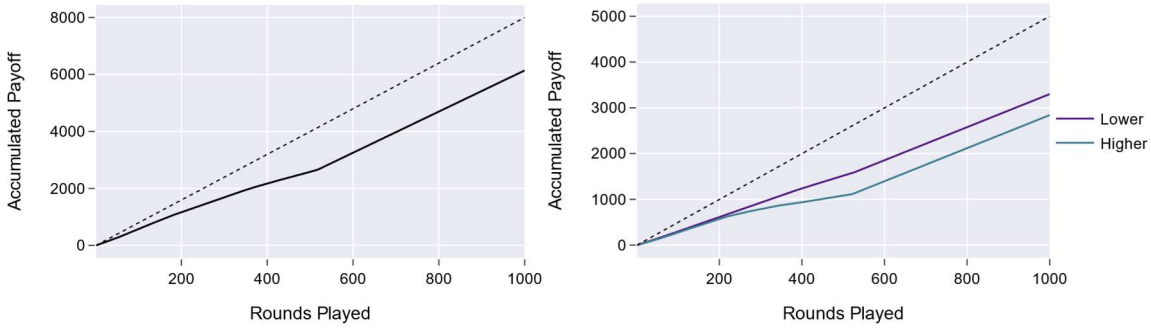

Figure 4.11c: (left) *Overall performance of the Q-learning agents during the 1000 rounds of the IPD game (thick line). The theoretically best performance is shown for comparison (dot-dashed line).* ***(right)*** *Performance of each Q-learning agent during the IPD game. The theoretically best performance for each agent is shown for comparison (dot-dashed line). Decreasing the T payoff and increasing the P and S payoffs. Ratio=0.1 (for all the payoffs), Interval=10.*

The 71% of CC states that was achieved using ratio 0.1 and interval 10 (Figure 4.11b) contrasts the result of 31.9% (Figure 4.5b) in which we used exactly the same configuration; 0.1 ratio and 10 rounds interval. It is also worth mentioning that a higher ratio and same interval gave 63.6% of CC states, 7 percentage points less, while a lower interval of 5 rounds and same 0.1 ratio produced even less CC states (55.9%). That indicates again that the interval plays a crucial role in determining the final result.

4.2.6 Summary and discussion on positive emotions

The most obvious way that enhanced self-control was the increment of the presence of positive emotions in a positive reinforcement way (increased R), which also revealed that a greater ratio magnitude and more frequent changes, enhanced it slightly more. The vital role of positive emotions in the exertion of self-control is highlighted by the psychologists (Robinson et al., 2013; Tice et al., 2004). The negative reinforcement methods that were about eliminating the presence of negative emotions (increased P, S separately) appeared to be as effective as the presence of positive emotions, irrespectively of the ratio and the interval parameters. However, the frequency of the change and the magnitude of the ratio did matter when we increased P and S at the same time. These results indicated that a frequent and subtle elimination of negative emotions *impair* self-control (31.9% CC states), a finding that confirms the motivating role of negative emotions such as anxiety and guilt, in self-control behavior (Loewenstein & O'Donoghue, 2006; Robinson et al., 2013). Next, when we decreased the T value with a small ratio and frequently (0.1, 10), which eases the internal cognitive conflict that the agents experience (Schacht & Sommer, 2012), also enhances dramatically the self-control behavior (70.6% CC states). We also tried to decrease the T value given a payoff matrix with higher initial T value (T=9). This higher levels of internal conflict scenario, reveal again the role of the number of rounds that elapse between the changes. That is, the negative emotions which the internal conflict elicits almost impair self-control when the decrement does not occur frequently enough or when the magnitude of the ratio is not large enough. Finally, explicitly eliminating the negative emotions (increased P and S) and easing the internal conflict (Cleanthous, 2010) by decreasing the Temptation payoff (T) at the same time, also achieved high levels of self-control.

4.3 Simulating negative emotions

After examining the different ways of simulating the presence of positive emotions and their effects on self-control, we now proceed in testing the methods of simulating the negative emotions and therefore the hypothesis that negative affect impairs self-control behavior. In order to test that we will change the payoff values (T, R, P, S) separately and

in various combinations during the 1000 rounds of the IPD game, according to a ratio and the number of rounds that will pass between each change, the interval. Again, we use the same initial payoff matrix (T=5, R=4, P=-2, S=-3), learning rates and epsilon values that were used in the constant payoff matrix experiment.

4.3.1 Decreasing the Reward payoff

The first method that we are going to use is to decrease the Reward (R) payoff since it would make it more tempting to defect. For this reason, we expect to see a decline in the overall CC states and ultimately to fail in exercising self-control. After testing various combinations of ratios and intervals, the lower percentage of CC states that we get is 49.9% (Figure 4.12a-b) when the ratio is set to 2 and the interval to 50 rounds and 54.6% with the same ratio but 10 rounds interval. This indicates that the CC states are slightly less when the negative emotions are experienced in greater intervals. Note that, the ratio cannot be set to values larger than 2 because then the IPD's second rule (2R > T+S) is violated.

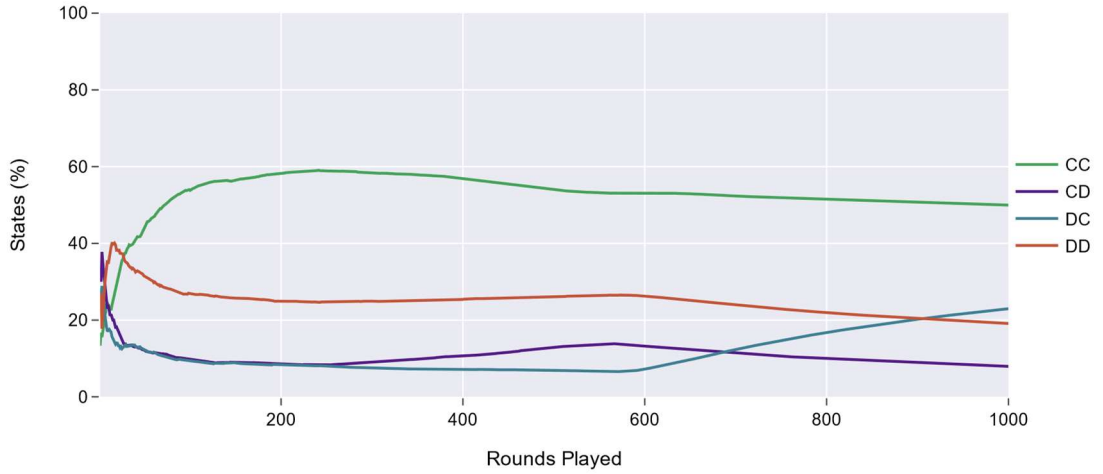

Figure 4.12a: *Average of the outcomes CC, CD, DC and DD during 1000 rounds of the Q-learning agents playing the IPD game. Decreasing the R payoff. Ratio=2, Interval=50.*

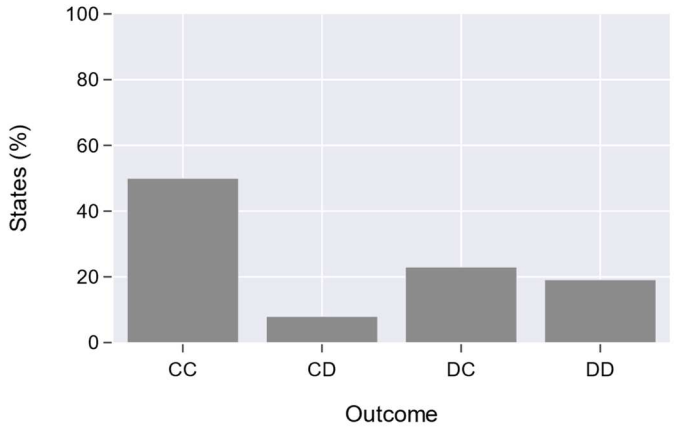

Figure 4.12b: *Overall average outcomes after 1000 rounds of the Q-learning agents playing the IPD game. Decreasing the R payoff. Ratio=2, Interval=50.*

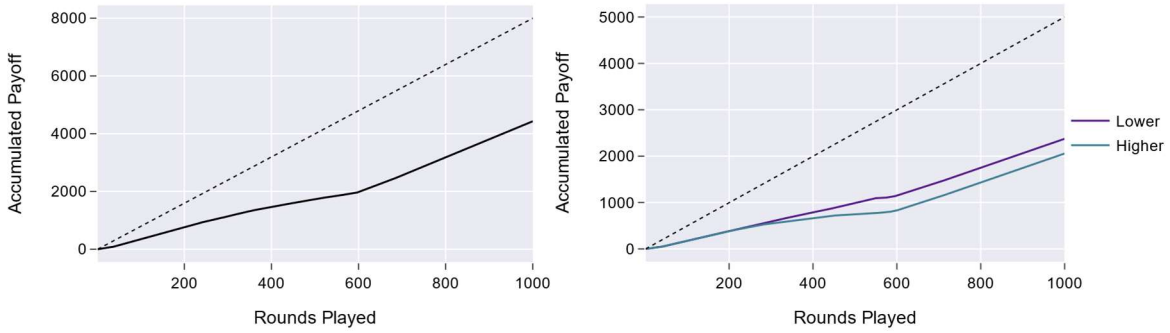

Figure 4.12c: (left) *Overall performance of the Q-learning agents during the 1000 rounds of the IPD game (thick line). The theoretically best performance is shown for comparison (dot-dashed line).* ***(right)*** *Performance of each Q-learning agent during the IPD game. The theoretically best performance for each agent is shown for comparison (dot-dashed line). Decreasing the R payoff. Ratio=2, Interval=50.*

Smaller ratios that were tested, not only did they not negatively affect self-control behavior, but also improved it. For example, the 0.1 ratio and 50 rounds interval resulted in 62.75% of CC states. Moreover, the 0.5 and 1 ratios with 10 rounds interval produced similar results. However, a 69.2% of CC states which approaches the good results that we got when testing the effects of positive emotions, was the result of a 0.2 ratio and 25 rounds interval. Figure 4.14 shows that result in comparison with the worst CC outcome that we produced when R was decreased by 2 every 50 rounds.

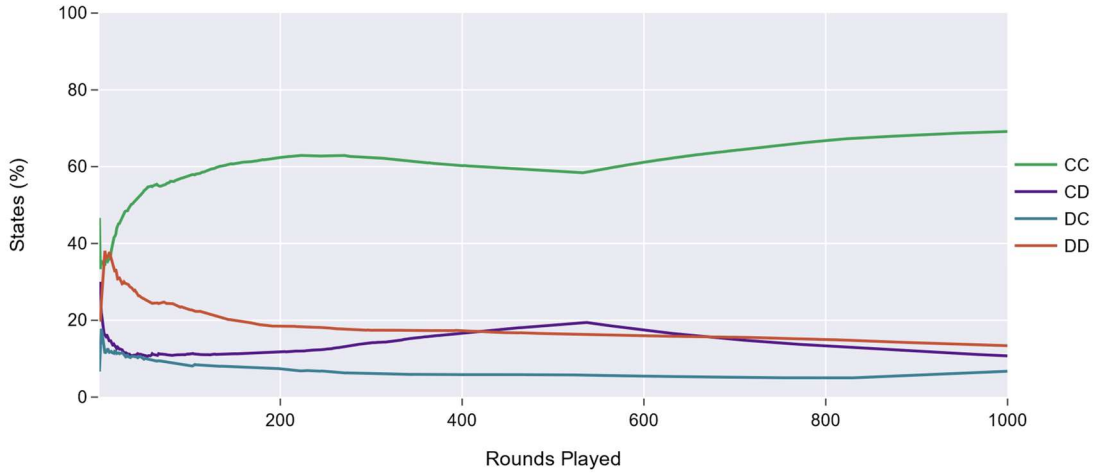

Figure 4.13: *Average of the outcomes CC, CD, DC and DD during 1000 rounds of the Q-learning agents playing the IPD game. Decreasing the R payoff. Ratio=0.2, Interval=25.*

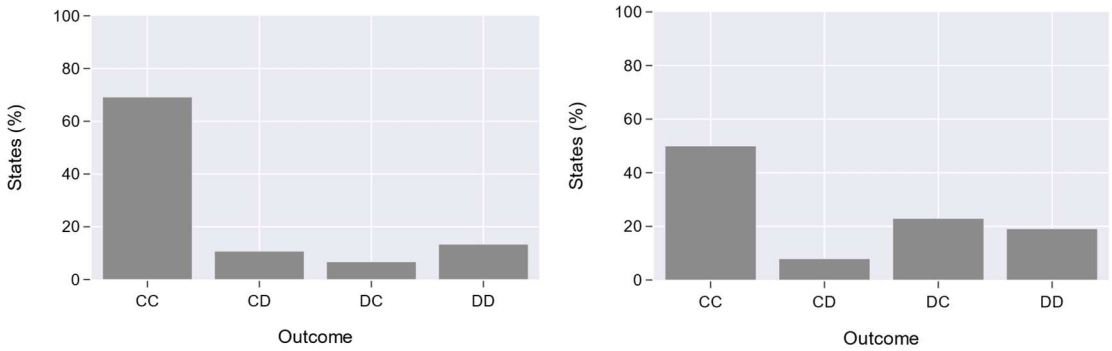

Figure 4.14: *Overall average outcomes after 1000 rounds of the Q-learning agents playing the IPD game.* ***(left)*** *Decreasing the R payoff. Ratio=0.25, Interval=25.* ***(right)*** *Decreasing the R payoff. Ratio=2, Interval=50.*

4.3.2 Increasing the Temptation payoff

The second method that was used is the increasement of the Temptation (T) payoff. The results after testing several combinations of ratios and intervals showed a decrement in the overall CC states, but the self-control was always achieved. The lower CC state outcome (55.7%) occurred when the T was increased by 5 every 250 rounds (Figures 4.15a, 4.15b). In fact, sometimes we got results slightly better than the baseline's — 64.9% of CC states was the result of ratio 2 and 50 rounds interval. Combinations of ratios and intervals smaller than 2 and 50 respectively produced similar to that. Therefore, what seems to have caused the decrement of the overall CC states is the large change (ratio>=5)

of the T value between large intervals (>=50). Also note that there is no point for the ratio to exceed the value of 10, since the second rule of the IPD game will not be satisfied (4*2=8 > 10-3=7) and the dilemma will no longer exist.

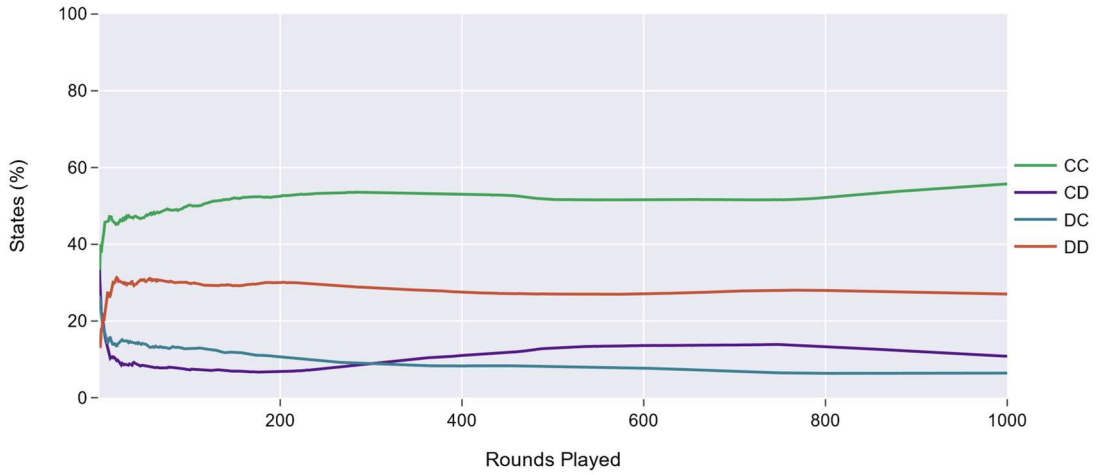

Figure 4.15a: *Average of the outcomes CC, CD, DC and DD during 1000 rounds of the Q-learning agents playing the IPD game. Increasing the T payoff. Ratio=5, Interval=250.*

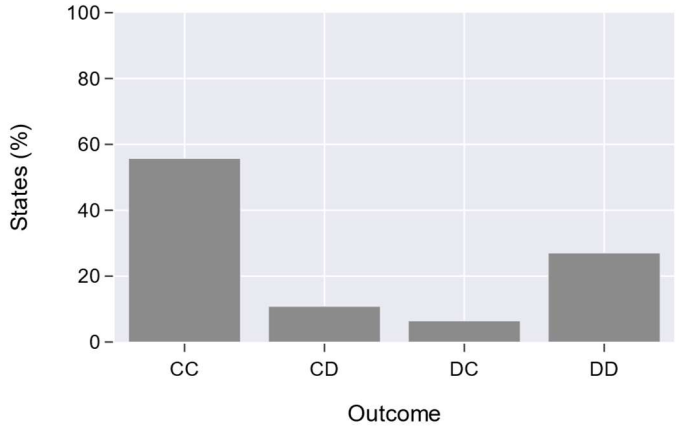

Figure 4.15b: *Average of the outcomes CC, CD, DC and DD during 1000 rounds of the Q-learning agents playing the IPD game. Increasing the T payoff. Ratio=5, Interval=250.*

4.3.3 Decreasing the Sucker's payoff

Next, by decreasing the Sucker's (S) payoff we provide the agents with even greater magnitudes of negative signals which represent the presence of negative emotional states. The ratio in which the S payoff will decrease is not limited by any rule of the IPD game. Hence ratios like 0.1, 0.5 and 1 were tested in combination with 10, 50 and 500 rounds

interval. The results showed that using any of that ratios and any intervals larger than 50, the CC states prevail, as the Figures 4.18 (left) show where a ratio 0.5 and interval 500 produced CC states at 53.8%. Figure 4.17a-b confirms that the self-control behavior is easily achieved under the last-mentioned configurations.

However, when a smaller interval of 10 rounds is used, along with 0.5 ratio, we experience a dramatic drop in the CC states (15.9%) while the DC states reach the 80.8% of the overall states. The DC state indicates the failure of the two agents to engage in cooperation since the agent that defects continues to get its positive reward (the Temptation payoff) and therefore it is not interested in cooperating. Moreover, because of the small interval of 10 rounds, the S payoff decreased significantly earlier in the game and thus, became impossible for the system to recover in contrast with larger intervals where the agents reached cooperation. Hence, it is obvious that when the magnitude of the negative signal is increased between small intervals, the self-control behavior is not achieved (Figure 4.16).

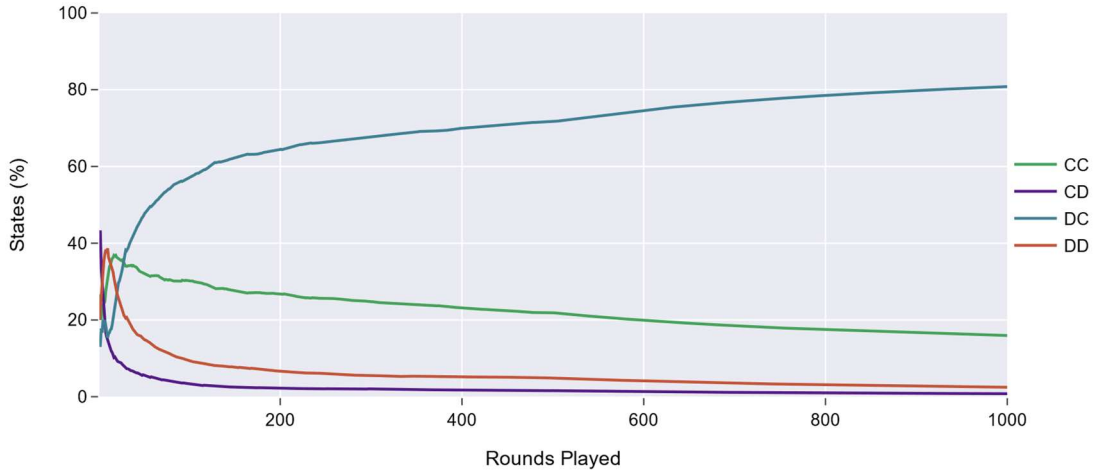

Figure 4.16: *Average of the outcomes CC, CD, DC and DD during 1000 rounds of the Q-learning agents playing the IPD game. Decreasing the S payoff. Ratio=0.5, Interval=10.*

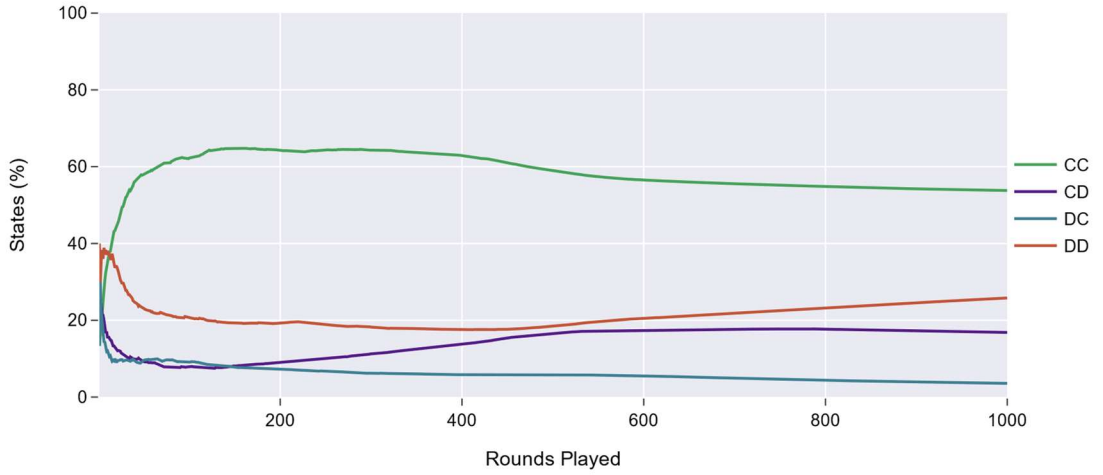

Figure 4.17a: *Average of the outcomes CC, CD, DC and DD during 1000 rounds of the Q-learning agents playing the IPD game. Decreasing the S payoff. Ratio=0.5, Interval=500.*

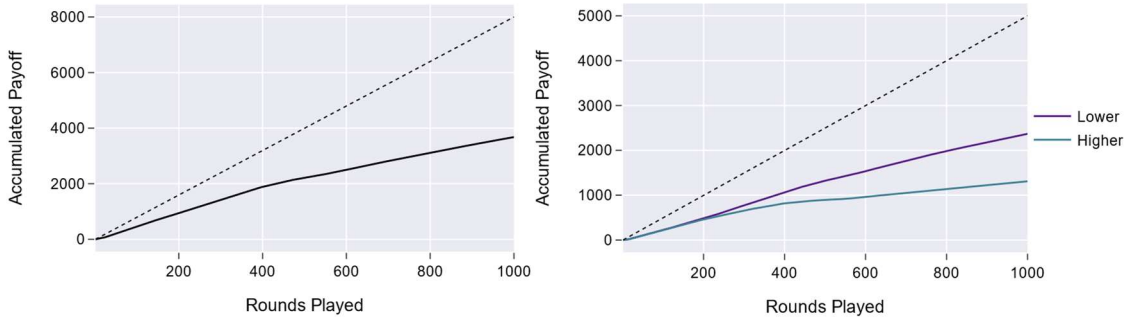

Figure 4.17b: (left) *Overall performance of the Q-learning agents during the 1000 rounds of the IPD game (thick line). The theoretically best performance is shown for comparison (dot-dashed line).* ***(right)*** *Performance of each Q-learning agent during the IPD game. The theoretically best performance for each agent is shown for comparison (dot-dashed line). Decreasing the S payoff. Ratio=0.5, Interval=500.*

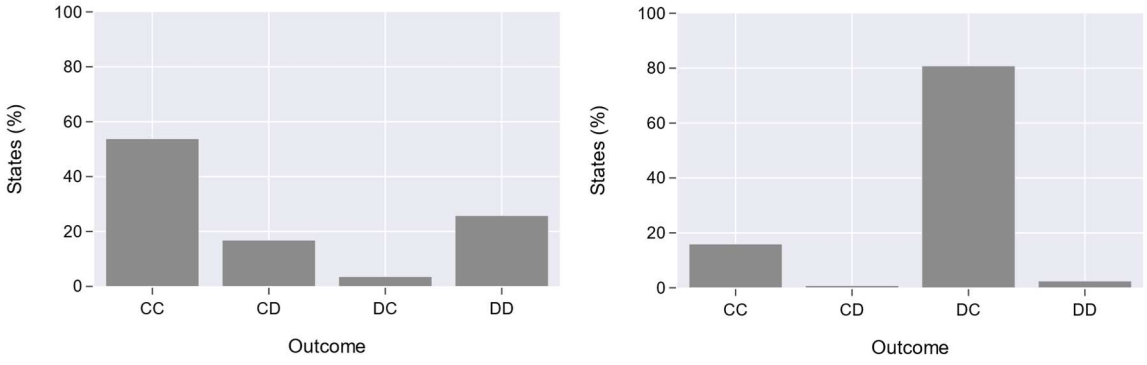

Figure 4.18: *Overall average outcomes after 1000 rounds of the Q-learning agents playing the IPD game.* ***(left)*** *Decreasing the S payoff. Ratio=0.5, Interval=500.* ***(right)*** *Decreasing the S payoff. Ratio=0.5, Interval=10.*

4.3.4 Increasing the Temptation and decreasing the Sucker's payoffs

This method is about increasing the Temptation (T) and decreasing the Sucker's (S) payoff at the same time. When we attempted in 4.3.2 to increase only the T payoff, there was not significant decrement of the CC states as we were expecting. However, by decreasing the S payoff in 4.3.3 we learned that only small intervals of 10 rounds lead to self-control failure. Consequently, we anticipate that the combination of the two methods will cause a further decline of CC states and thus, self-control failure.

As was expected, the DC state prevailed when any ratio was combined (0.2, 0.5, 1 for both payoffs) with small intervals of 10 rounds (Figure 4.19). When the interval was increased to 50 rounds, we had to also increase the magnitude of the ratio in order to produce the same effect. Figure 4.20 shows that ratio 1 and interval 50 still results in self-control failure but just after the first 200 rounds of the game. However, the use of a large interval (≥50) results in successful self-control behavior when it is combined with a small ratio (≤0.5), as Figure 4.21 shows.

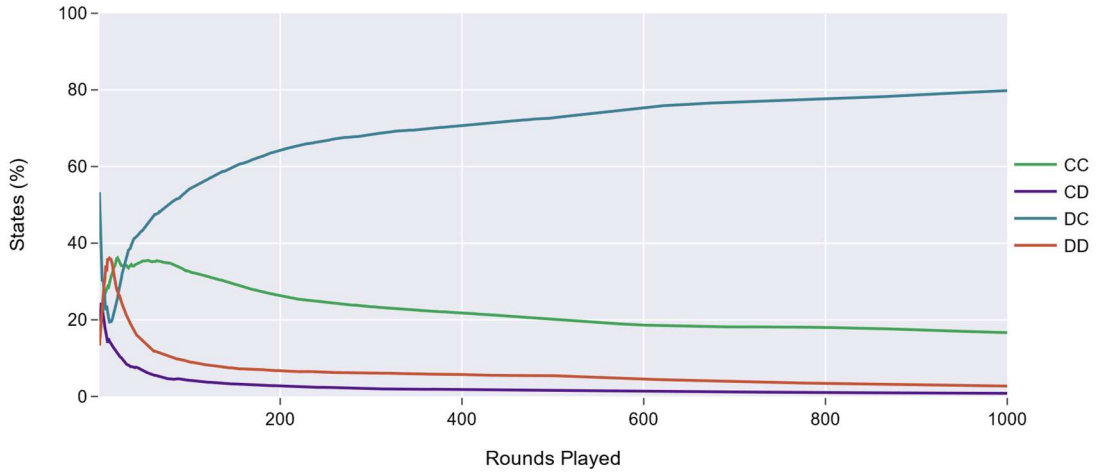

Figure 4.19: *Average of the outcomes CC, CD, DC and DD during 1000 rounds of the Q-learning agents playing the IPD game. Increasing the T payoff and the decreasing the S payoff. Ratio=0.5 (for both payoffs), Interval=10.*

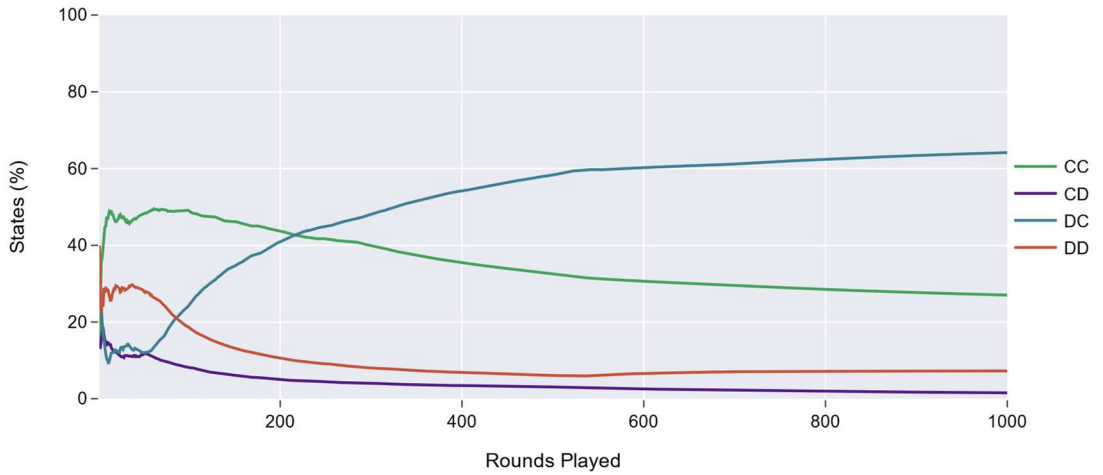

Figure 4.20: *Average of the outcomes CC, CD, DC and DD during 1000 rounds of the Q-learning agents playing the IPD game. Increasing the T payoff and the decreasing the S payoff. Ratio=1 (for both payoffs), Interval=50.*

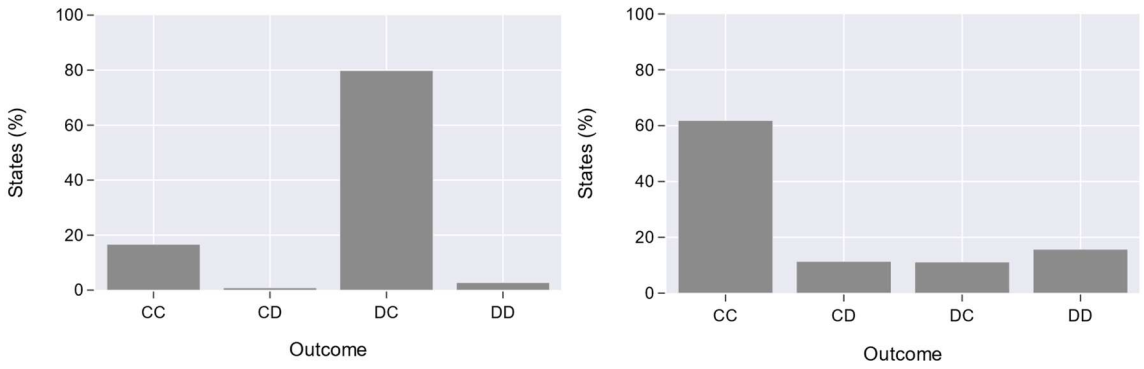

Figure 4.21: *Overall average outcomes after 1000 rounds of the Q-learning agents playing the IPD game.* ***(left)*** *Increasing the T payoff and the decreasing the S payoff. Ratio=0.5 (for both), Interval=10.* ***(right)*** *Increasing the T payoff and the decreasing the S payoff. Ratio=0.5 (for both), Interval=50.*

4.3.5 Decreasing the Punishment payoff

Figures 4.22a and 4.22b show that the method of decreasing the Punishment (P) payoff resulted in self-control behavior. The overall CC states reached 57.2%, while the overall DD states reached 21.7%, one of the highest rates of DD states. Since the P payoff has only 1-point difference with the Sucker's payoff, no large ratios (>0.5) are useful, since we would break the game's first rule after the second change of the payoff's value. Although that kind of ratios (>0.5) were tested, the results were similar with the ones

produced with ratios like 0.1 and 0.5 and intervals varying from 50 to 500 rounds. Figure 4.22c shows the low accumulated payoff which also indicates the decrement in the self-control levels.

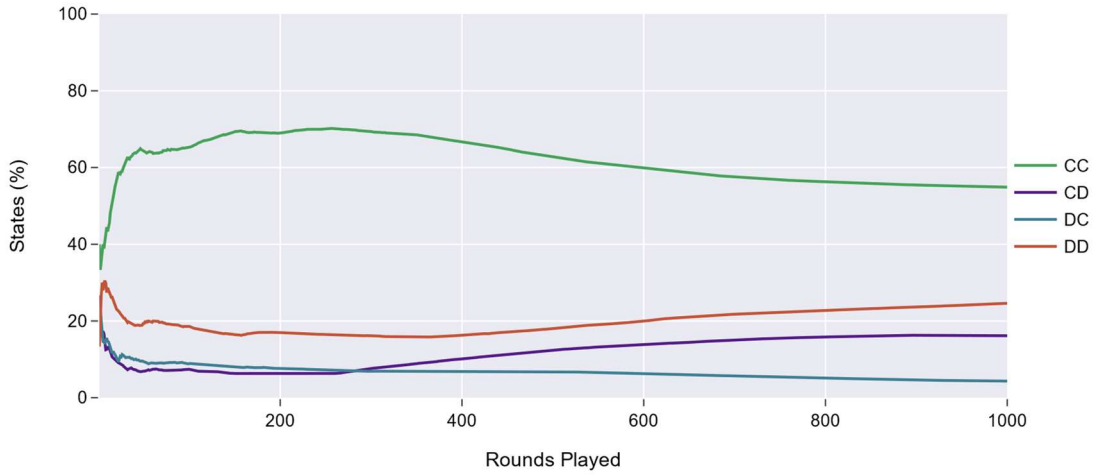

Figure 4.22a: *Average of the outcomes CC, CD, DC and DD during 1000 rounds of the Q-learning agents playing the IPD game. Decreasing the P payoff. Ratio=0.1, Interval=50.*

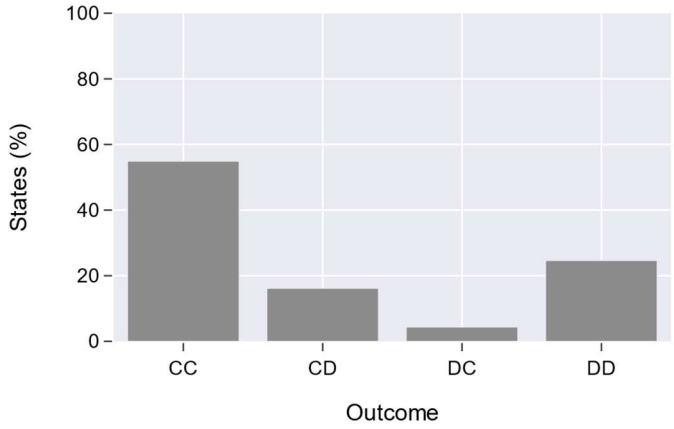

Figure 4.22b: *Overall average outcomes after 1000 rounds of the Q-learning agents playing the IPD game. Decreasing the P payoff. Ratio=0.1, Interval=50.*

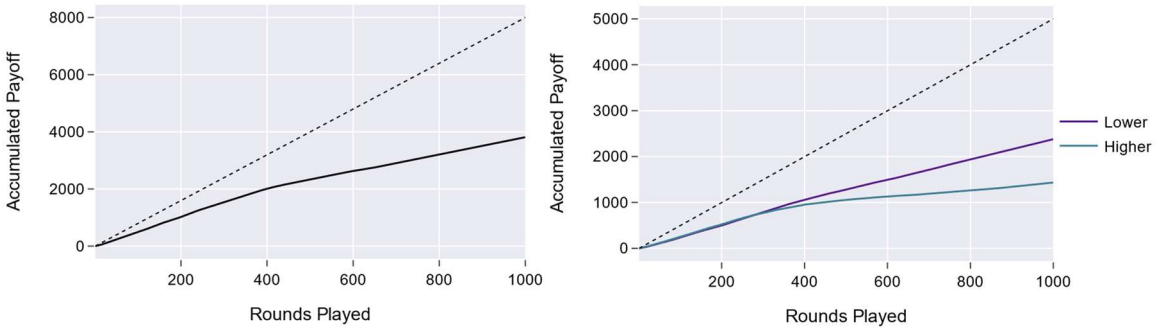

Figure 4.22c: (left) *Overall performance of the Q-learning agents during the 1000 rounds of the IPD game (thick line). The theoretically best performance is shown for comparison (dot-dashed line).* ***(right)*** *Performance of each Q-learning agent during the IPD game. The theoretically best performance for each agent is shown for comparison (dot-dashed line). Decreasing the P payoff. Ratio=0.1, Interval=50.*

4.3.6 Increasing the Temptation and decreasing the Punishment payoffs

When it was attempted to increase the Temptation (T) and the Punishment (P) payoff separately (sections 4.3.2 and 4.3.5), we noticed that the no matter the magnitude of the ratio and the interval, the self-control behavior was achieved. Combining these two methods gives similar results. Experimenting showed that regardless the value of the ratios and the interval, the agents engaged in self-control behavior with the CC states ranging from 50.9% to 61.6%. Figures 4.24a and 4.24b in which ratios 0.1 and interval 5 are used, show this effect.

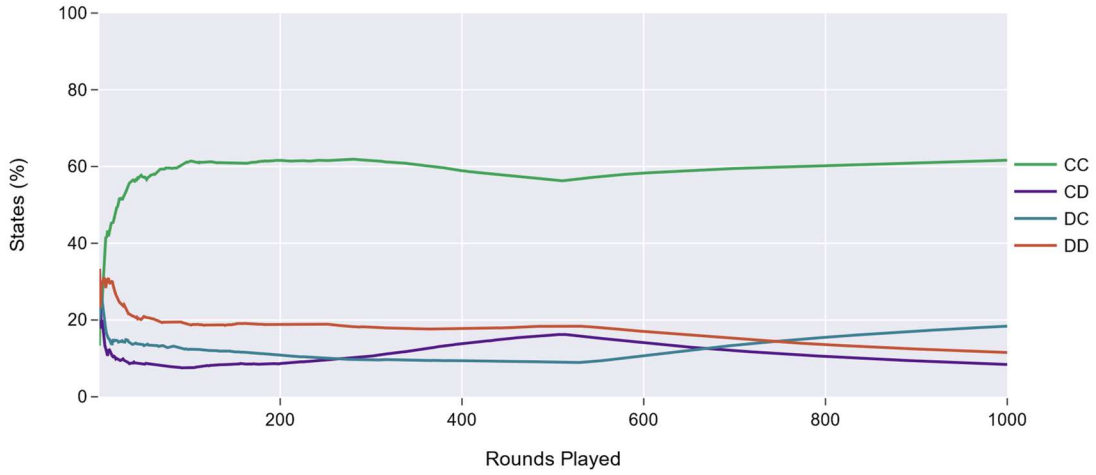

Figure 4.24a: *Average of the outcomes CC, CD, DC and DD during 1000 rounds of the Q-learning agents playing the IPD game. Increasing the T payoff and decreasing the P payoff. Ratio=0.1 (for both), Interval=5.*

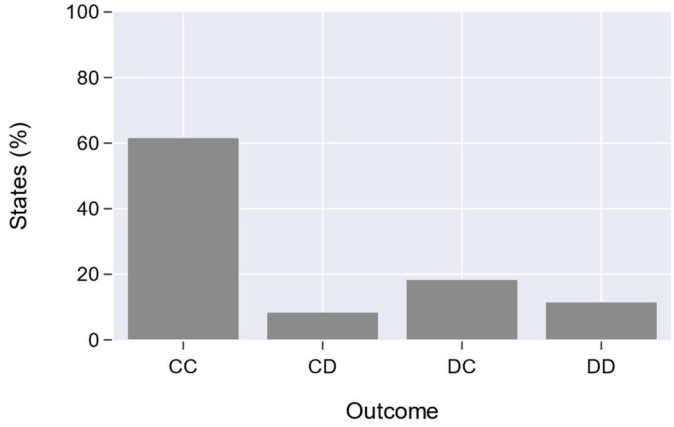

Figure 4.24b: *Overall average outcomes after 1000 rounds of the Q-learning agents playing the IPD game. Increasing the T payoff and decreasing the P payoff. Ratio=0.1 (for both), Interval=5.*

4.3.7 Decreasing the Punishment and the Sucker's payoffs

Another combination of methods that produced interesting results was decreasing the Punishment (P) and Sucker's (S) payoffs at the same time. When we were examining the effects of decreasing only the S payoff, we concluded that small intervals result in self-control failure. Indeed, as Figures 4.25 and 4.28 (right) shows, when the interval was set to 10 rounds, the P ratio to 0.5 and the S ratio to 1, the CC states dropped at 15.9% and the DC states reached 80.8% of all states.

It was rather unexpected that when the interval was kept at 10 rounds and both ratios were set to 0.5, the CC states increased to 51.8% and self-control was 'just' achieved (Figure 4.25). The change influenced the decrement of the S ratio (from 1 to 0.5), a result that we did not get when we tested the 0.5 ratio and interval 10 in the method of only decreasing S (section 4.3.2). So, it seems that the self-control failure in Figure 4.26 was caused by the increased magnitude of the S ratio.

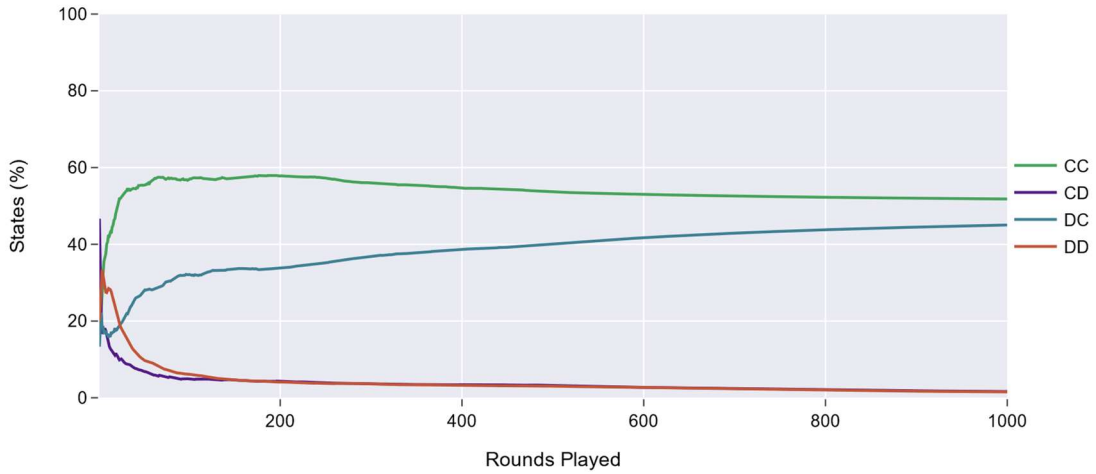

Figure 4.25: *Average of the outcomes CC, CD, DC and DD during 1000 rounds of the Q-learning agents playing the IPD game. Decreasing the P and the S payoff. Ratio=0.5 (for both), Interval=10.*

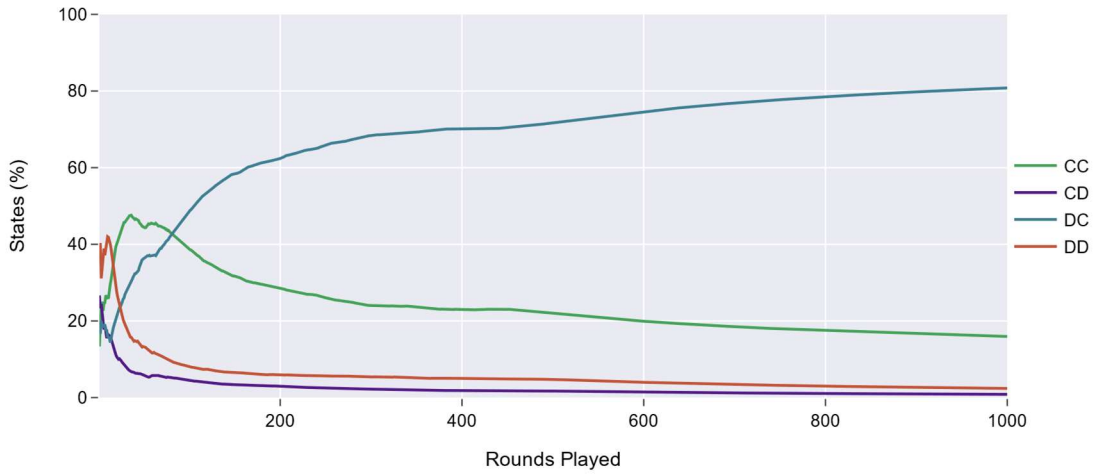

Figure 4.26: *Average of the outcomes CC, CD, DC and DD during 1000 rounds of the Q-learning agents playing the IPD game. Decreasing the P and the S payoff. Ratio P=0.5, Ratio S=1, Interval=10*

After achieving self-control with both ratios at 0.5 and interval 10 rounds, we tested the effect of the same ratios but interval 50 rounds. The CC states reached 70.9% (Figures 4.27, 4.28 *left*), a significantly high value, since it was not easily encountered even in the subsections of 4.2 where we were simulating the presence of positive emotions. Figure 4.28 presents in comparison the two configurations (0.5, 0.1, 10 vs 0.5, 0.5, 50). Furthermore, when the P and S were decreased separately, self-control was achieved but never improved in comparison with our baseline's results. This result will help us later to

support the case that is found in the literature, that negative emotions do improve self-control.

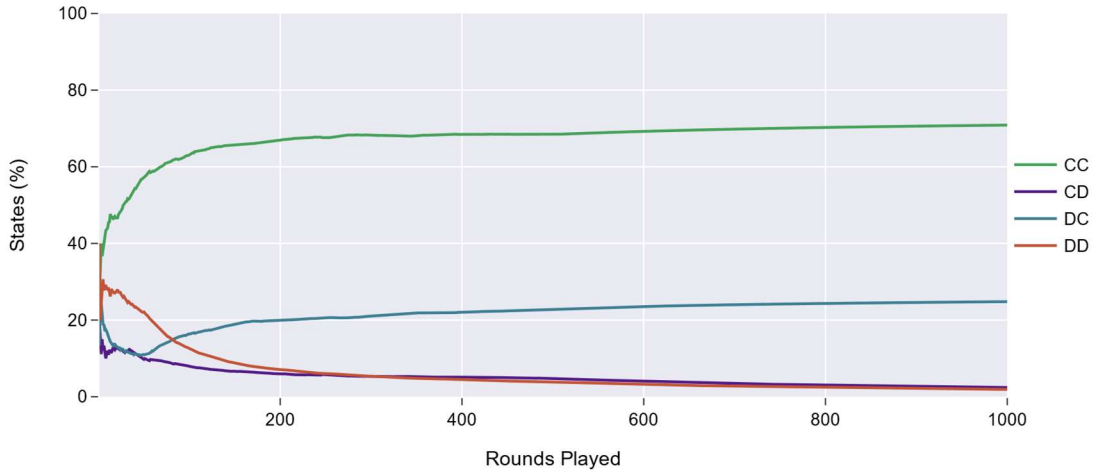

Figure 4.27: *Average of the outcomes CC, CD, DC and DD during 1000 rounds of the Q-learning agents playing the IPD game. Decreasing the P and the S payoff. Ratio=0.5 (for both), Interval=50.*

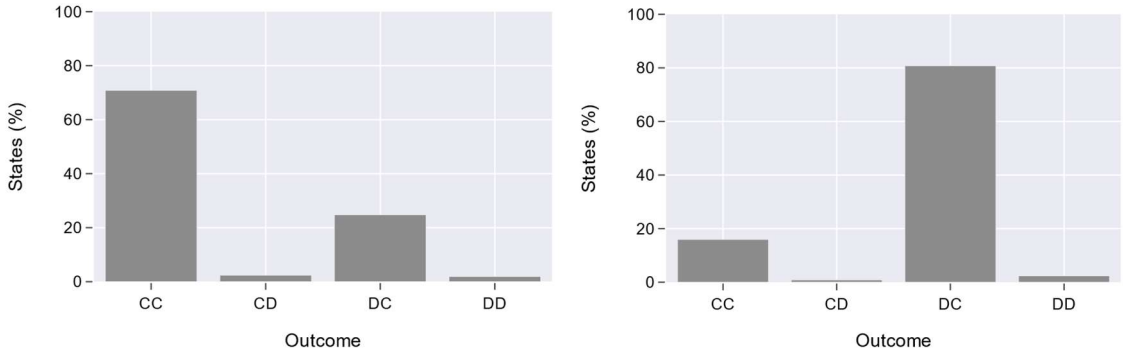

Figure 4.28: *Overall average outcomes after 1000 rounds of the Q-learning agents playing the IPD game.* ***(left)*** *Decreasing the P and the S payoff. Ratio=0.5 (for both), Interval=50.* ***(right)*** *Decreasing the P and the S payoff. Ratio P=0.5, Ratio T=1, Interval=10*

4.3.8 Decreasing the Reward and the Punishment payoffs

The combination of decreasing the Reward (R) and the Punishment (P) payoffs at the same time, results in the same levels of self-control as in sections 4.3.2 and 4.3.5 where the two methods were tested separately. Figures 4.29a and 4.29b are the result of P ratio 0.1, R ratio 0.5 and 50 rounds interval and represent the results of tests with a variety of combinations. Even the sudden decrease of R (with ratio 2) that produced the lowest

levels of CC states, had not the same effect in combination with decreasing the P payoff. Moreover, the combination that achieved remarkable high CC states results when we only decreased R (0.2, 25 in Figure 4.13), had not the same effect in combination with decreasing the P payoff. The decrement ratios of the R and P payoffs are restricted by the game's first rule.

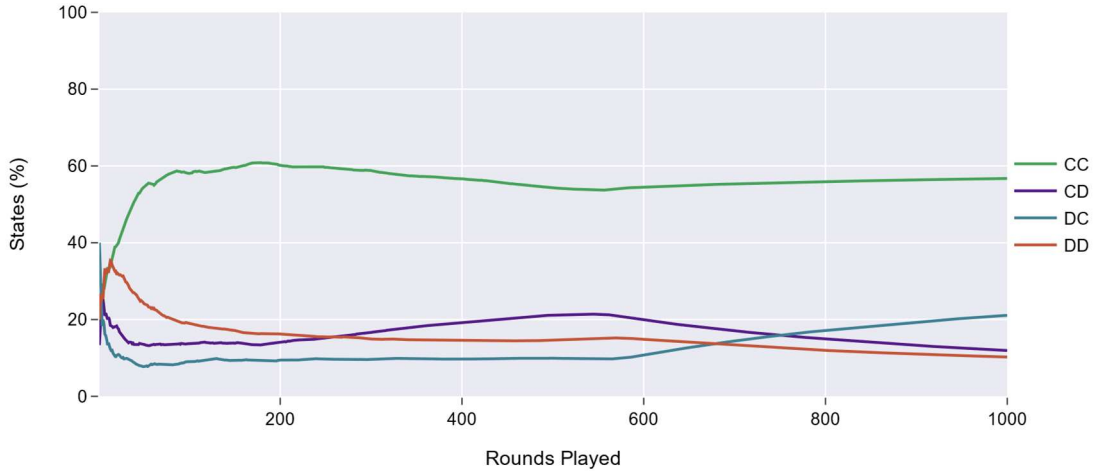

Figure 4.29a: *Average of the outcomes CC, CD, DC and DD during 1000 rounds of the Q-learning agents playing the IPD game. Decreasing the R and the P payoffs. Ratio R=0.1, Ratio P=0.5, Interval=50.*

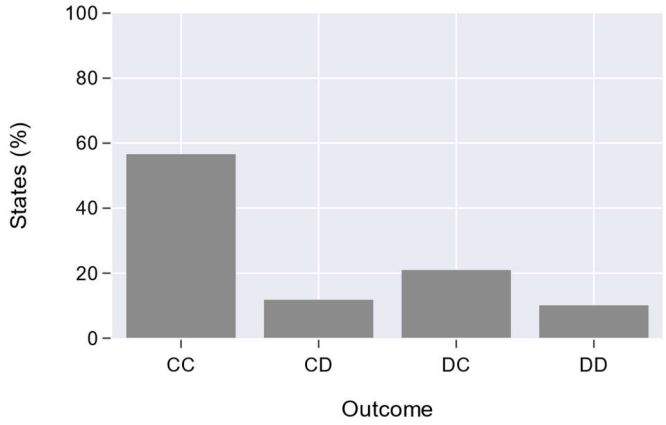

Figure 4.29b: *Overall average outcomes after 1000 rounds of the Q-learning agents playing the IPD game. Decreasing the R and the P payoffs. Ratio R=0.1, Ratio P=0.5, Interval=50.*

4.3.9 Decreasing the Reward and the Sucker's payoffs

This time we combine the decrement of the Reward (R) and the Sucker's payoff (S). When only the R payoff was decreased (section 4.3.1), self-control was always achieved, in contrast with the method of only decreasing the S payoff (section 4.3.3) which resulted in self-control failure when the interval was too small (<10). Hence, we now expect to see a further decrement of the CC states.

With R ratio 0.1, S ratio 1 and interval 50 rounds, DC states occurred at 53.4% and CC states at 37.4% (Figure 4.30) and thus resulting in self-control failure. One might think that since the method of only decreasing R did not result in failure, what influenced the result in this case is the decrement of S. However, when the method of only decreasing S, was tested with S ratio at 1 and interval 50, self-control was achieved (55.1% CC states) as we can see in comparison in Figure 4.31. Therefore, it is clear that the decrement of the R payoff played a crucial role.

Moreover, when the S ratio is decreased from 1 to 0.5, self-control is again achieved (Figure 4.32) — as happened when we decreased only the S payoff with 0.5 ratio and a large interval of 50 rounds (section 4.3.3). So, decreasing the R payoff has not affect this outcome. Similarly, the same ratio that produce the lowest CC states when we decreased only the R payoff (ratio 2, interval 50), in combination now with ratio 1 for the T payoff, has not produced any further decrement.

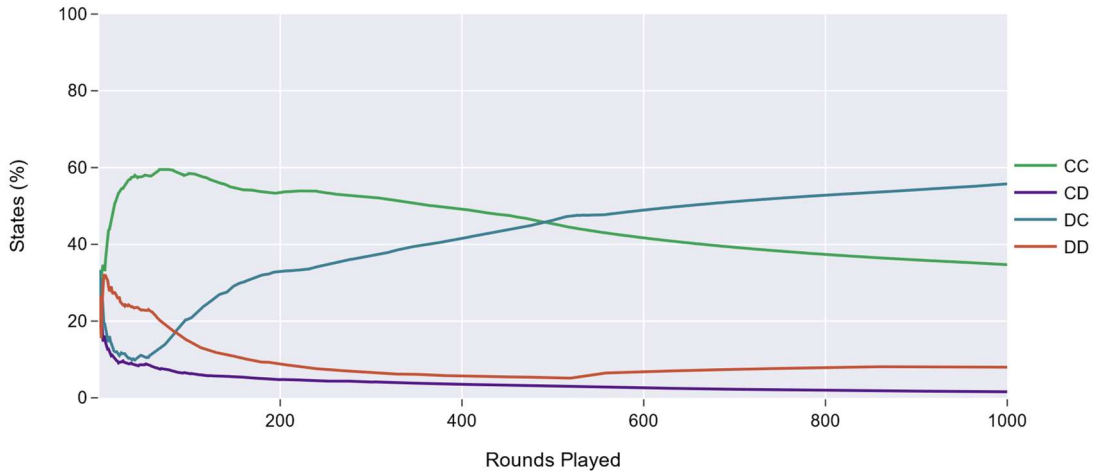

Figure 4.30: *Average of the outcomes CC, CD, DC and DD during 1000 rounds of the Q-learning agents playing the IPD game. Decreasing the R and the S payoffs. Ratio R=0.1, Ratio S=1, Interval=50.*

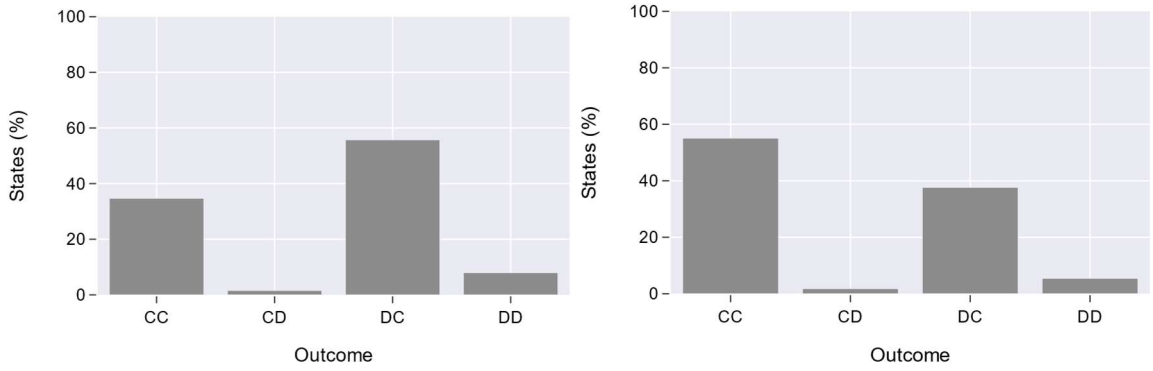

Figure 4.31: *Overall average outcomes after 1000 rounds of the Q-learning agents playing the IPD game.* ***(left)*** *Decreasing the R and the S payoffs. Ratio R=0.1, Ratio S=1, Interval=50.* ***(right)*** *Decreasing the S payoff. Ratio=1, Interval=50.*

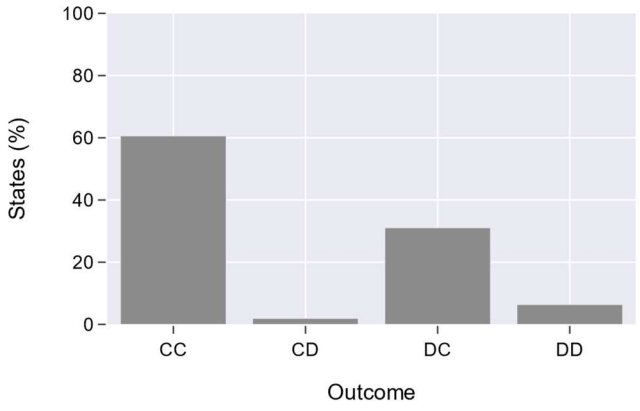

Figure 4.32: *Overall average outcomes after 1000 rounds of the Q-learning agents playing the IPD game. Decreasing the R and the S payoffs. Ratio R=0.1, Ratio S=0.5, Interval=50*

.

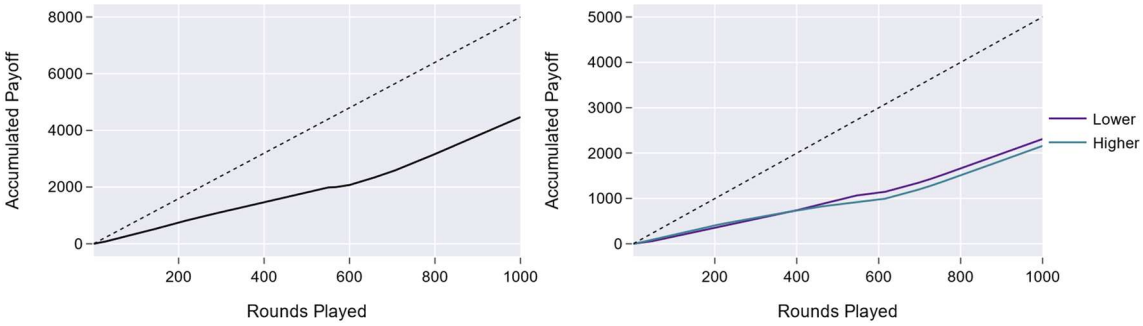

Figure 4.23c: (left) *Overall performance of the Q-learning agents during the 1000 rounds of the IPD game (thick line). The theoretically best performance is shown for comparison (dot-dashed line).* ***(right)*** *Performance of each Q-learning agent during the IPD game. The theoretically best performance for each agent is shown for comparison (dot-dashed line). Decreasing the R and the S payoff. Ratio R=2, Ratio S=1, Interval=50.*

4.3.10 Decreasing the Reward, Punishment and Sucker's payoffs

By decreasing the Sucker's payoff (S) at the same time with the Reward (R) and Sucker's payoffs (S), we expect to see a further reduction of the CC states and an increment of the DC states. After all, when only the R and the P payoffs were decreased, we obtained results close to the baselines. In addition, the P ratio could not be set to large values due to the small difference with the S payoff and the potential violation of the IPD's first rule.

We set the R ratio to 0.1 and the S ratio to 1, since that values effectively influenced the result in section 4.3.9. Figures 4.33 and 4.34 (left) reveal the self-control failure when the aforementioned configuration was set, along with the P ratio at 0.5 and 50 rounds interval. The overall CC states reached only 40.4%, 15 percentage points less than when only the R and P payoffs were decreased (Figure 4.29b). Nevertheless, the decrement of the R and the S payoffs only, appears to impair self-control to a further extent, since with the same configurations, the DC states reached at 37.4% of the overall states (Figure 4.31 *left*).

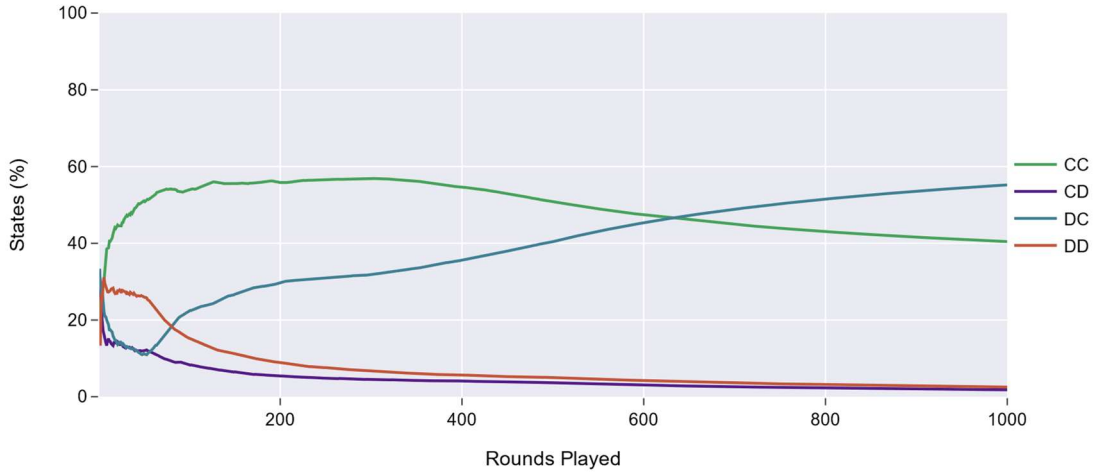

Figure 4.33: *Average of the outcomes CC, CD, DC and DD during 1000 rounds of the Q-learning agents playing the IPD game. Decreasing the R, the P and the S payoffs. Ratio R=0.1, Ratio P=0.5, Ratio S=1, Interval=50.*

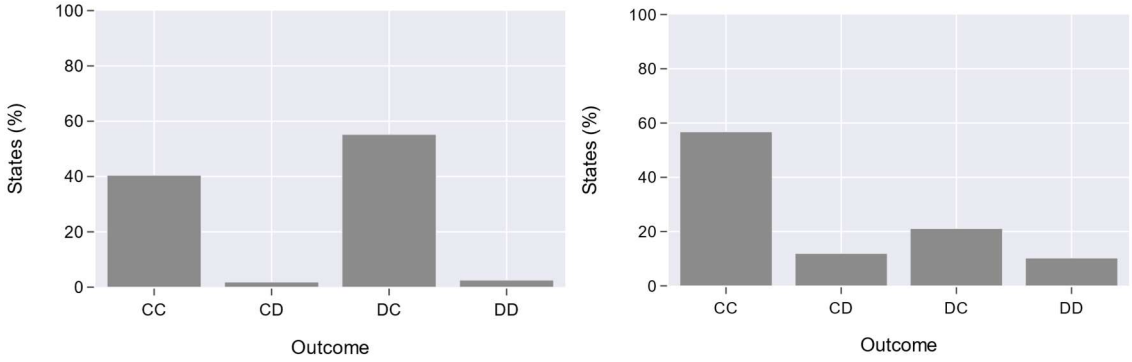

Figure 4.34: *Overall average outcomes after 1000 rounds of the Q-learning agents playing the IPD game.* ***(left)*** *Decreasing the R, the P and the S payoffs. Ratio R=0.1, Ratio P=0.5, Ratio S=1, Interval=50* ***(right)*** *Decreasing the R and the P payoffs. Ratio R=0.1, Ratio P=0.5, Interval=50*

4.3.11 Summary and discussion on negative emotions

First, we test a method of negative punishment (decreased R) which simulates the elimination of positive emotions and found out that self-control is moderated when the ratio's increment/decrement is large and the change infrequent. In the contrary, when the ratio is small and frequent, the self-control behavior is enhanced (69.2% CC states). Next, the agents seem to always learn to cooperate when we were increasing the T value alone (increment of *positive* emotions in each agent, but increment of *negative* emotions for the system), or at the same time with decreasing the P value. Note that the decrement of the

P value also did not cause self-control failure but reduced obviously the levels of self-control and increased the frequency of the DD states (24.6%). This reveals how the agents prefer the suboptimal choice of defecting more often one another in the presence of new P values in small intervals. The DD outcome is after all the only Nash equilibrium (Nash, 1950), so it seems that the introduction of new P values and the changed environment dynamics, guides them to insist to their choice instead of cooperating. When the S value was decreased in small intervals, the DC states reached 80.8% of the overall states, indicating clearly the agents' failure to engage in cooperation. Since the agent who defects continues to get its positive reward, there is no reason to switch its tactic to cooperation. However, when the change is subtle, self-control is still achieved but in lower levels (53.8%). Not surprisingly then, increasing the T and decreasing the S value at the same time and frequently, caused self-control failure (80% DC states). When the increment of negative emotions is not frequent, the agents cooperate successfully. Until now, the explicit increment of negative emotions (S, P decrement separately) and with combination with the negative emotions elicited by the increased internal conflict (T increment) (Schacht & Sommer, 2012) caused low levels, or complete failure of self-control. This is in compliance with the psychological literature about the negative effects of negative emotions on self-control (Chester et al., 2016; Cyders & Smith, 2008; Tice et al., 2004). Moreover, we found evidence supporting the necessity of negative emotions in successful self-control behavior (Giner-Sorolla, 2001; Loewenstein & O'Donoghue, 2006). The decrement of S and P at the same time infrequently produced high levels of self-control (70.9% CC states). However, when the decrement was occurring more often, the agents failed in cooperation —another time that the number of rounds between the changes matters (Rolls, 2012). Lastly, the combination of eliminating the positive emotions (decreased R) with increasing the presence of negative ones (decreased P), resulted in lower levels of self-control (55% CC states), and even in self-control failure (40.4% CC states) when all of the methods were deployed simultaneously (decreased R, P, S). It is noteworthy that the concurrent elimination of positive emotions *did* caused the self-control failure (37.4%) in a configuration that only presence of negative emotions did not (decreased R and S).

Chapter 5

Conclusions and Future Work

5.1 Overview and conclusions

The current thesis examined the effects of positive and negative emotions on a computational model of self-control. Rachlin (2000) provided us with a self-control definition that is also supported by the neuroscientists that investigate the mechanisms of self-control in the brain; self-control is the competition between the higher part of the brain that is responsible for cognition, and the lowest part which is responsible for motivation. The need to simulate that interaction, lead to deploying a general-sum game, the Prisoner's Dilemma game. The higher and lower parts of the brain engage in a variation of the PD game, the Iterated Prisoner's Dilemma (IPD) game, and their goal is to cooperate , a state that represents the achieved self-control. Meanwhile, we decided to use Rolls' (2012) definition of emotions as "states elicited by instrumental reinforcers" in order to use the values of the payoff matrix of the IPD game and represent the effects that the presence of positive and negative emotions have on the self-control behavior. The values of the payoff matrix are the instrumental reinforcers, since the same values are the rewards that the agents receive and learn to cooperate. Therefore, emotional states are elicited and the effects of them are depicted on the results (whether the agents learnt or not to cooperate and the overall accumulated payoff).

Again, Rolls' (2012) definition of emotions helped us define them as positive or negative, according to whether the reinforcer was positive or negative. The effect of positive and negative emotions has been studied first by psychologists and later by neuroscientists.

The first suggestion was that positive emotions promote self-control, whereas negative ones impair it (Baumeister, 2004). By using fMRI techniques, it was proven that the excessive deployment of the prefrontal cortex and continuous self-control failure leads to ego depletion. However, it has been also suggested that experiencing emotions intensely (even positive ones), leads to not focusing on the long-term, but on the immediate rewards. On the other hand, negative emotions sometimes help to exercise self-control by transferring the negative consequences of not reaching the long-term goal, to the present.

Like Georgiou (2015), we used the Q-Learning algorithm to train the two agents in learning mutual cooperation, while one or more values of the payoff matrix were changing throughout the rounds of the game to simulate the presence of emotions, and without violating the two fundamental rules of the IPD game. Beginning every simulation with the same payoff matrix and learning rates, we were differentiating two important factors: the ratios of the payoff values that would change, that is a positive or negative value that was added on the payoff values, and the number of rounds that would elapse between every change of the payoff values. The agents were experiencing positive emotions when the positive value R was increasing (explicitly positive emotions, a kind of positive reinforcement), or when the positive value T was decreasing (negative reinforcement), or when the negative values P and S were increasing (elimination of negative emotion and thus, improved mood). The agents were experiencing negative emotions when the T value was increasing (positive punishment), or when the R was decreasing (elimination of positive emotions, a kind of negative punishment), or when the P and S were increasing (explicit negative emotions).

After running the simulation with different combinations of ratios and intervals (in rounds), we found out that our model experiences the effects of the positive and negative emotions as they are suggested by the Psychology and Neuroscience literature. Regarding the effects of the positive emotions, first we achieved to improve self-control when we increased their intensity by increasing the Reward payoff (Robinson et al., 2013; Tice et al., 2004). In addition, the elimination of the negative emotions infrequently as well as the ease of the internal conflict (decreased T) were methods that maintained or enhanced self-control levels. In fact, we achieved the highest level of self-control (71% CC states) when we were simultaneously decreasing the Temptation payoff and increasing the

Punishment and Sucker's payoffs, which suggests that the elimination of negative emotions caused by the internal cognitive conflict (Schacht & Sommer, 2012) is as effective as the increment of positive ones. Secondly, we found out that extreme elimination of the negative emotions leads to self-control failure (P and S increasing frequently). This result demonstrates the effectiveness of the totally necessary negative emotions, like guilt and fear in the self-regulatory processes (Giner-Sorolla, 2001; Loewenstein & O'Donoghue, 2006).

The effects of the negative emotions on the behavior of the Q-Learning agents are the following. The method of the negative punishment (decreased R) decreased the levels of self-control, only when negative emotion was given rarely and in large doses. Moreover, the agents have achieved self-control despite the presence of explicit negative emotions elicited by increased internal conflict (increased T) (Schacht & Sommer, 2012). Similarly, explicit increment of negative emotions (decreased P) did not caused self-control failure, but an increased frequency of DD states was observed, which indicates that the frequently changing dynamics, forced the agents to insist on defecting.

On the contrary, self-control totally failed when the S payoff value was decreased frequently, resulting in the record of 80.8% DC states. This phenomenon was not noticed when the S payoff was decreasing in a much greater interval (50 to 500 rounds), suggesting that the more frequent the change, the more profound the impact, a finding which correlates with the neuroscience studies on the *continuous* self-control failure in the presence of negative emotions (Chester et al., 2016; VanderVeen et al., 2016). Until now, the explicit increment of negative emotions (S, P decrement separately) and with combination with the negative emotions elicited by the increased internal conflict (T increment) (Schacht & Sommer, 2012) caused low levels, or complete failure of self-control. This is in compliance with the psychological literature about the negative effects of negative emotions on self-control (Chester et al., 2016; Cyders & Smith, 2008; Tice et al., 2004).

It is noteworthy that the concurrent elimination of positive emotions *did* cause the self-control failure (37.4%) in a configuration that only the presence of negative emotions did not (decreased R and S). Here, the fact that one might be able to exert self-control even

under negative emotions is highlighted, but once the positive emotions are eliminated too, then exerting self-control is not possible anymore. We also decreased P at the same time with S and we proved the case that negative emotions *do* help in self-control (Giner-Sorolla, 2001; Loewenstein & O'Donoghue, 2006). High levels of self-control were achieved (70.9% CC states), which the agents were unable to maintain once the change was taking place more frequently. Self-control failure was caused though, when a small interval and high in magnitude ratios were given (80% DC states), proving for one more time the importance of the number of rounds that elapsed between each change (Rolls, 2012).

In conclusion, we managed to incorporate the concept of emotions in a computational model of self-control, and simulate not the emotions *per se*, but rather the consequences of their presence. These effects were in compliance with the psychological theories on self-control and the applied research of neuroscientists —in other words, the model is cognitively adequate. Furthermore, these simulations confirmed once again, the structure of the self-control and the appropriateness of the PD game for modeling it. Equally important is that the flexibility of the model was demonstrated, since it allowed the addition of the dimensions of the intensity (increment/decrement of the reinforcing signals) of the emotions and the interval of change. Understanding the self-control behavior is about understanding human nature. Thus, any contribution towards that direction is worthwhile.

5.2 Future Work

We examined the effect of positive and negative emotions on self-control. Some of the effects enhanced this behavior, while others appear to have devastating results on it. The goal of Behavioral psychologists is to help overcome self-control failures by suggesting several techniques with the most prominent to be based on the strength model of self-control (Baumeister et al., 2007), which was tested by Cleanthous (2010) who was interchanging constant payoff matrices of low and high internal conflict. This approach could also be tested using our Q-Learning model with non-constant payoff matrices this time. Emotion regulation is another relevant research interface that the Psychological literature provides, which suggests new ways of improving self-control when emotions

impair it. Learning how self-control processes intervene with emotion regulation and vice versa has been the goal of many theoretical and empirical studies (Gross, 1999; Koole et al., 2016), thus testing those insights that one can apply in order to improve self-control, could be the next step of this work.

Another direction this work could follow is adding an explainability layer on top of the self-control computational model which would have the role of the "mind over the brain". This means that the explainability layer's ("mind") goal is to interpret in a systematic way the decisions made by the model underneath ("brain") by using a framework called Cognitive Argumentation suggested by Saldanha & Kakas (2020) and relies on the assumption that the task of modeling human behavior and reasoning is computational and implementable. That is, we have already provided the implementation and we have linked the results to the findings of the Cognitive Sciences (Psychology and Neuroscience) which explain them. Connecting the results to their cognitive interpretation is what inspired this new perspective. Part of this task is to also cognitively interpret or *cognitively validate* the two rules (T>R>P>S, 2R>T+S) which are necessary conditions for the IPD. Working towards this direction will contribute to the general attempt to model the human behavior.

We have started the discussion on self-control based on the premise that humans act as rational agents whose goal is to achieve the optimum behavior (Stanovich, 2012). However, this is not the only definition of the term "rational". Besides the normative definition of rationality, rationality is descriptively defined as any decision that is followed by a statistically significant proportion of the population (Saldanha & Kakas, 2020). So, another goal of developing the explainability layer on top of the self-control model, would be to manage to interpret behaviors based on this extra rationality definition. After all, positioning the underlying premises on rationality, which ultimately explain why we consider the CC state as the desired state and the goal of self-control, in the Great Rationality Debate (Stanovich, 2011), will only help to comprehend self-control more accurately.

References

Banfield, G. D. (2006). *Simulation of Self-Control through Precommitment Behaviour in an Evolutionary System*. Ph.D. University of London.

Barrett, L. F. (2017). *How Emotions Are Made: The Secret Life of the Brain*. Boston, MA: Houghton Mifflin Harcourt.

Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1994). Guilt: An interpersonal approach. *Psychological Bulletin*, *115*(2), 243–267.

Baumeister, Roy F., Vohs, K. D., & Tice, D. M. (2007). The Strength Model of Self-Control. *Current Directions in Psychological Science*, *16*(6), 351–355.

Berman, M. G., Yourganov, G., Askren, M. K., Ayduk, O., Casey, B. J., Gotlib, I. H., Kross, E., McIntosh, A. R., Strother, S., Wilson, N. L., Zayas, V., Mischel, W., Shoda, Y., & Jonides, J. (2013). Dimensionality of brain networks linked to life-long individual differences in self-control. *Nature Communications*, *4*(1), 1373.

Bremner, J. D., Staib, L. H., Kaloupek, D., Southwick, S. M., Soufer, R., & Charney, D. S. (1999). Neural correlates of exposure to traumatic pictures and sound in Vietnam combat veterans with and without posttraumatic stress disorder: A positron emission tomography study. *Biological Psychiatry*, *45*(7), 806–816.

Calkins, S. D., & Howse, R. B. (2004). Individual differences in self-regulation: Implications for childhood adjustment. In P. Philippot & R. S. Feldman (Eds.), *The regulation of emotion* (pp. 307–332). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Chester, D. S., Lynam, D. R., Milich, R., Powell, D. K., Andersen, A. H., & DeWall, C. N. (2016). How do negative emotions impair self-control? A neural model of negative urgency. *NeuroImage*, *132*, 43–50.

Christodoulou, C., Banfield, G., & Cleanthous, A. (2010). Self-control with spiking and non-spiking neural networks playing games. *Journal of Physiology, Paris*, *104*(3–4), 108–117.

Cleanthous, A. (2010). *In search of self-control through computational modelling of internal conflict*. Ph.D. University of Cyprus.

Curtis, C. E., & D'Esposito, M. (2003). Success and failure suppressing reflexive behavior. *Journal of Cognitive Neuroscience*, *15*(3), 409–418.

Cyders, M. A., & Smith, G. T. (2008). Emotion-based Dispositions to Rash Action: Positive and Negative Urgency. *Psychological Bulletin*, *134*(6), 807–828.

Cyders, M. A., Smith, G. T., Spillane, N. S., Fischer, S., Annus, A. M., & Peterson, C. (2007). Integration of impulsivity and positive mood to predict risky behavior: Development and validation of a measure of positive urgency. *Psychological Assessment*, *19*(1), 107–118.

Dreisbach, G. (2006). How positive affect modulates cognitive control: The costs and benefits of reduced maintenance capability. *Brain and Cognition*, *60*(1), 11–19.

Duckworth, A. L., & Seligman, M. E. P. (2005). Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychological Science*, *16*(12), 939–944.

Fischer, S., Smith, G. T., Spillane, N. S., & Cyders, M. A. (2005). Urgency: Individual Differences in Reaction to Mood and Implications for Addictive Behaviors. In *Psychology of moods* (pp. 85–107). Hauppauge, NY: Nova Science Publishers.

Georgiou, A. (2015). *Μελέτη σχέσης αυτοελέγχου και συνειδητότητας*. Bachelor's thesis. University of Cyprus.

Giner-Sorolla, R. (2001). Guilty pleasures and grim necessities: Affective attitudes in dilemmas of self-control. *Journal of Personality and Social Psychology*, *80*(2), 206–221.

Goldberg, E. (2002). *The Executive Brain: Frontal Lobes and the Civilized Mind*. Oxford, England: Oxford University Press.

Gross, J. J. (1999). Emotion Regulation: Past, Present, Future. *Cognition and Emotion*, *13*(5), 551–573.

Hajcak, G., Molnar, C., George, M. S., Bolger, K., Koola, J., & Nahas, Z. (2007). Emotion facilitates action: A transcranial magnetic stimulation study of motor cortex excitability during picture viewing. *Psychophysiology*, *44*(1), 91–97.

Hare, T. A., Camerer, C. F., & Rangel, A. (2009). Self-control in decision-making involves modulation of the vmPFC valuation system. *Science (New York)*, *324*(5927), 646–648.

Heatherton, T. F. (2011). Neuroscience of Self and Self-Regulation. *Annual Review of Psychology*, *62*, 363–390.

Herrnstein, R. J. (1990). Rational choice theory: Necessary but not sufficient. *American Psychologist*, *45*(3), 356–367.

Holub, A., Hodgins, D. C., & Peden, N. E. (2005). Development of the temptations for gambling questionnaire: A measure of temptation in recently quit gamblers. *Addiction Research & Theory*, *13*(2), 179–191.

Isen, A. M. (1987). Positive Affect, Cognitive Processes, and Social Behavior. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 20, pp. 203–253). Cambridge, MA: Academic Press.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kavka, G. S. (1991). Is Individual Choice Less Problematic than Collective Choice? *Economics & Philosophy*, *7*(2), 143–165.

Kirby, K. N., & Herrnstein, R. J. (1995). Preference Reversals Due to Myopic Discounting of Delayed Reward: *Psychological Science*, *6*(2), 83–89.

Knoch, D., & Fehr, E. (2007). Resisting the Power of Temptations. *Annals of the New York Academy of Sciences*, *1104*(1), 123–134.

Knutson, B., & Greer, S. M. (2008). Anticipatory affect: Neural correlates and consequences for choice. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1511), 3771–3786.

Koole, S. L., van Dille, L. F., & Sheppes, G. (2016). The Self-Regulation of Emotion. In K. D. Vohs & R. F. Baumeister (Eds.), *Handbook of Self-Regulation: Theoretical and Empirical Advances* (3rd ed., pp. 101–112). New York: Guilford.

Kraut, R. (2018). Aristotle's Ethics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (2018th ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2018/entries/aristotle-ethics/

Loewenstein, G., & O'Donoghue, T. (2006). 'We Can Do This the Easy Way or the Hard Way': Negative Emotions, Self-Regulation, and the Law. *The University of Chicago Law Review*, *73*(1), 183–206.

MacIntyre, P. D., & Vincze, L. (2017). Positive and negative emotions in motivation for second language learning. *Studies in Second Language Learning and Teaching*, *7*(1), 61–88.

McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science*, *306*(5695), 503–507.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202.

Mischel, W., Shoda, Y., & Rodriguez, M. I. (1989). Delay of gratification in children. *Science*, *244*(4907), 933–938.

Mischel, Walter, & Ebbesen, E. B. (1970). Attention in delay of gratification. *Journal of Personality and Social Psychology*, *16*(2), 329–337.

Nash, J. F. (1950). Equilibrium Points in n-Person Games. *Proceedings of the National Academy of Sciences of the United States of America*, *36*(1), 48–49.

Nygren, T. E., Isen, A. M., Taylor, P. J., & Dulin, J. (1996). The Influence of Positive Affect on the Decision Rule in Risk Situations: Focus on Outcome (and Especially Avoidance of Loss) Rather Than Probability. *Organizational Behavior and Human Decision Processes*, *66*(1), 59–72.

Osborne, M. J. (2004). *An Introduction to Game Theory* (Vol. 3). Oxford, England: Oxford University Press.

Rachlin, H. (1995). Self-control: Beyond commitment. *Behavioral and Brain Sciences*, *18*(1), 109–121.

Rachlin, H. (2000). *The Science of Self-Control*. Harvard University Press.

Rapoport, A., Chammah, A. M., & Orwant, C. J. (1965). *Prisoner's Dilemma: A Study in Conflict and Cooperation*. Ann Arbor, MI: University of Michigan Press.

Reeve, J. (2014). *Understanding Motivation and Emotion*. John Wiley & Sons.

Reisenzein, R. (2017). The Legacy of Cognition-Arousal Theory: Introduction to a Special Section of Emotion Review. *Emotion Review*, *9*(1), 3–6.

Ren, J., Hu, L., Zhang, H., & Huang, Z. (2010). Implicit Positive Emotion Counteracts Ego Depletion. *Social Behavior and Personality: An International Journal*, *38*(7), 919–928.

Robertson, S., Davies, M., & Winefield, H. (2017). Positive psychological correlates of successful weight maintenance in Australia. *Clinical Psychologist*, *21*(3), 236–244.

Robinson, M. D., Watkins, E. R., & Harmon-Jones, E. (2013). *Handbook of Cognition and Emotion*. Hoboken, NJ: Guilford Press.

Rolls, E. T. (2012). *Neuroculture: On the implications of brain science*. Oxford, England: Oxford University Press.

Rolls, E. T. (2013). What are Emotional States, and Why Do We Have Them? *Emotion Review*, *5*(3), 241–247.

Rubinstein, A. (1982). Perfect Equilibrium in a Bargaining Model. *Econometrica*, *50*(1), 97–109.

Ruderman, A. J. (1985). Dysphoric mood and overeating: A test of restraint theory's disinhibition hypothesis. *Journal of Abnormal Psychology*, *94*(1), 78–85.

Rummery, G. A., & Niranjan, M. (1994). *On-Line Q-Learning Using Connectionist Systems* (CUED/F-INFENG/TR No. 166). Department of Engineering, Cambridge University.

Ryan, R. M., & Deci, E. L. (2001). On Happiness and Human Potentials: A Review of Research on Hedonic and Eudaimonic Well-Being. *Annual Review of Psychology*, *52*(1), 141–166.

Saldanha, E.-A. D., & Kakas, A. (2020). Cognitive Argumentation and the Suppression Task. *ArXiv Preprint ArXiv:2002.10149*.

Schacht, A., & Sommer, W. (2012). Emotions in Cognitive Conflicts. In N. M. Seel (Ed.), *Encyclopedia of the Sciences of Learning* (pp. 1139–1141). Boston MA: Springer US.

Schmeichel, B. J., & Tang, D. (2015). Individual Differences in Executive Functioning and Their Relationship to Emotional Processes and Responses. *Current Directions in Psychological Science*, *24*(2), 93–98.

Skinner, B. F. (1938). *The Behavior of Organisms: An Experimental Analysis*. New York: Appleton-Century-Crofts.

Stanovich, K. E. (2011). *Rationality and the Reflective Mind*. Oxford, England: Oxford University Press.

Stanovich, K. E. (2012). On the Distinction Between Rationality and Intelligence: Implications for Understanding Individual Differences in Reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford Handbook of Thinking and Reasoning* (pp. 433–455). Oxford, England: Oxford University Press.

Stucke, T. S., & Baumeister, R. F. (2006). Ego depletion and aggressive behavior: Is the inhibition of aggression a limited resource? *European Journal of Social Psychology*, *36*(1), 1–13.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, *3*(1), 9–44.

Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High Self-Control Predicts Good Adjustment, Less Pathology, Better Grades, and Interpersonal Success. *Journal of Personality*, *72*(2), 271–324.

Tice, D. M., Baumeister, R. F., & Zhang, L. (2004). The Role of Emotion in Self-Regulation: Differing Roles of Positive and Negative Emotion. In P. Philippot & R. S. Feldman (Eds.), *The Regulation of Emotion* (pp. 215–230). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

van Otterlo, M., & Wiering, M. (2012). Reinforcement Learning and Markov Decision Processes. In M. Wiering & M. van Otterlo (Eds.), *Reinforcement Learning: State-of-the-Art* (pp. 3–42). Berlin, Heidelberg: Springer Berlin Heidelberg.

VanderVeen, J. D., Plawecki, M. H., Millward, J. B., Hays, J., Kareken, D. A., O'Connor, S., & Cyders, M. A. (2016). Negative urgency, mood induction, and alcohol seeking behaviors. *Drug and Alcohol Dependence*, *165*, 151–158.

Watkins, C. J. C. H. (1989). *Learning from delayed rewards.* Ph.D. University of Cambridge.

Wörgötter, F., & Porr, B. (2008). Reinforcement learning. *Scholarpedia*, *3*, 1448.

Appendix A

The development of the basic Q-Learning model (Appendices A-B) is by Georgiou (2015).

Class Main.java

```
package com.example.QlearningModel;

import java.io.BufferedWriter;
import java.io.File;
import java.io.FileWriter;
import java.io.IOException;

/**
 * Created by jeannettechahwan & annageorgiou on 20/01/15.
 */
public class Main {
   public static Player lower;
   public static Player higher;

   public static double[][][] states_results;
   public static double[][][] payoff_results;
   public static double[] overall_payoff;

   public static double T, R, P, S;

   public static int get_state(int action_lower, int action_higher) {
      int next_state;

      if(action_lower==0 && action_higher==0)
         next_state = 0;
      else if(action_lower==0 && action_higher==1)
         next_state = 2;
      else if(action_lower==1 && action_higher==0)
         next_state = 1;
      else
         next_state = 3;

      return next_state;
   }

   public static void update_states_results(int episode, int state, int trial){
      states_results[episode][trial][0] = states_results[episode-1][trial][0];
      states_results[episode][trial][1] = states_results[episode-1][trial][1];
      states_results[episode][trial][2] = states_results[episode-1][trial][2];
      states_results[episode][trial][3] = states_results[episode-1][trial][3];
      states_results[episode][trial][state]++;
   }

   public static void update_payoff_results(int episode, double lower_payoff, double
higher_payoff, int trial){
      payoff_results[episode][trial][0] = payoff_results[episode-1][trial][0];
      payoff_results[episode][trial][1] = payoff_results[episode-1][trial][1];
```

```
    payoff_results[episode][trial][0] += lower_payoff;
    payoff_results[episode][trial][1] += higher_payoff;
  }

  public static void statistics(int numberOfTrials, int numberOfEpisodes){
    int e, t, s, p;
    double sumAll_states;
    double sumAll_payoffs;

    for(e=0; e<numberOfEpisodes; e++){
      for(t=0; t<numberOfTrials; t++){   // find sum of states visits from each trial of current
episode and save in last column
        for(s=0; s<4; s++){    // states
          states_results[e][numberOfTrials][s] += states_results[e][t][s];
        }

        for(p=0; p<2; p++){    // payoffs
          payoff_results[e][numberOfTrials][p] += payoff_results[e][t][p];
          payoff_results_neutral[e][numberOfTrials][p] += payoff_results_neutral[e][t][p];
        }
      }

      sumAll_states=0;
      sumAll_payoffs=0;

      for(s=0; s<4; s++){    // sum of state fields in last column (#trials+1)
        sumAll_states += states_results[e][numberOfTrials][s];
      }

      for(s=0; s<4; s++){    // normalise
        states_results[e][numberOfTrials][s] /= sumAll_states;
      }

      for(p=0; p<2; p++){    // sum of payoff fields in last column (#trials+1)
        sumAll_payoffs += payoff_results[e][numberOfTrials][p];
      }

      overall_payoff[e] = sumAll_payoffs/numberOfTrials;
    }

  }

  public static void print_statistics(int numberOfTrials, int numberOfEpisodes){
    int e, s, p;

    // states
    for(e=0; e<numberOfEpisodes; e++){
      for(s=0; s<4; s++){
        System.out.printf("%.4f ", states_results[e][numberOfTrials][s]);
      }
      System.out.println();
    }

    // payoffs
    for(e=0; e<numberOfEpisodes; e++){
      for(p=0; p<2; p++){
        System.out.printf("%.4f ", payoff_results[e][numberOfTrials][p]);
```

```
        }
        System.out.println();
    }
}

public static void printToFile_statistics(int numberOfTrials, int numberOfEpisodes){
    int e, s, p;

    // write results to file
    try {
        File file_s = new File("states_results.txt");
        File file_p = new File("payoff_results.txt");

        // if file doesnt exists, then create it
        if (!file_s.exists()) {
            file_s.createNewFile();
        }

        if (!file_p.exists()) {
            file_p.createNewFile();
        }

        FileWriter fw_s = new FileWriter(file_s.getAbsoluteFile());
        FileWriter fw_p = new FileWriter(file_p.getAbsoluteFile());
        BufferedWriter bw_s = new BufferedWriter(fw_s);
        BufferedWriter bw_p = new BufferedWriter(fw_p);

        for(e=0; e<numberOfEpisodes; e++){
            for(s=0; s<4; s++){
                bw_s.write(Double.toString(states_results[e][numberOfTrials][s]));
                bw_s.write(" ");
            }
            bw_s.write("\n");

            for(p=0; p<2; p++){
                bw_p.write(Double.toString(payoff_results[e][numberOfTrials][p]/numberOfTrials));
                bw_p.write(" ");
            }
            bw_p.write(Double.toString(overall_payoff[e]));
            bw_p.write(" ");
            bw_p.write(Double.toString(overall_payoff_neutral[e]));
            bw_p.write("\n");
        }

        bw_s.close();
        bw_p.close();
    } catch (IOException ex) {
        ex.printStackTrace();
    }
}

public static void main(String[] args) {
    int initial_state, current_state, next_state; // Takes values 0-3
    int i;
    int episodes_before = 500, episodes_after = 500;
    int trials = 15;
    states_results = new double[episodes_before + episodes_after][trials+1][4];
    payoff_results = new double[episodes_before + episodes_after][trials+1][2];
```

```
    overall_payoff = new double[episodes_before + episodes_after];

    payoff_results_neutral = new double[episodes_before + episodes_after][trials+1][2];
    overall_payoff_neutral = new double[episodes_before + episodes_after];

    T=5; R=4; P=-2; S=-3;

    int rounds = 25;
    double ratio_R = -0.2, ratio_T = 0;
    double ratio_P = -0.1, ratio_S = 0;

    for(int t=0; t<trials; t++) {
      lower = new Player();
      lower.setDiscount(0.1);
      lower.setLearning_rate(0.1);
      lower.setEpsilon(0.1);
      lower.setPayoff_matrix(4, 5, -3, -2);

      higher = new Player();
      higher.setDiscount(0.9);
      higher.setLearning_rate(0.1);
      higher.setEpsilon(0.1);
      higher.setPayoff_matrix(4, -3, 5, -2);

      initial_state = (int) (Math.random() * 4);
      current_state = initial_state;
      states_results[0][t][current_state]++;

      for (i = 1; i < episodes_before; i++) {
        lower.setCurrent_action(lower.choose_action(current_state));
        higher.setCurrent_action(higher.choose_action(current_state));

        next_state = get_state(lower.getCurrent_action(), higher.getCurrent_action());

        lower.update_Q(current_state, next_state);
        higher.update_Q(current_state, next_state);

        if ( i%rounds==0 &&
            ((T+ratio_T) > (R+ratio_R)) &&
            ((R+ratio_R) > (P+ratio_P)) &&
            ((P+ratio_P) > (S+ratio_S)) &&
            (2*(R+ratio_R) > ((T+ratio_T) + (S+ratio_S)))
          ){
          R = R + ratio_R;
          T = T + ratio_T;
          S = S + ratio_S;
          P = P + ratio_P;
          higher.setPayoff_matrix(R, S, T, P);
          lower.setPayoff_matrix(R, S, T, P);
          System.out.printf("R:%.4f T:%.4f S:%.4f P:%.4f e:%d t:%d\n",
higher.getPayoff_matrix(0),
              higher.getPayoff_matrix(2), higher.getPayoff_matrix(1),
              higher.getPayoff_matrix(3), i, t);
        }

        current_state = next_state;

        update_states_results(i, current_state, t);
```

```
            update_payoff_results(i, lower.getPayoff_matrix(current_state),
higher.getPayoff_matrix(current_state), t);
        }

        // Set epsilon = 0 and initial payoff matrix
        lower.setEpsilon(0);
        higher.setEpsilon(0);
        R=4; T=5; S=-3; P=-2;
        lower.setPayoff_matrix(4, 5, -3, -2);
        higher.setPayoff_matrix(4, -3, 5, -2);

        for (i = 1; i <= episodes_after; i++) {
          lower.setCurrent_action(lower.choose_action(current_state));
          higher.setCurrent_action(higher.choose_action(current_state));

          next_state = get_state(lower.getCurrent_action(), higher.getCurrent_action());

          lower.update_Q(current_state, next_state);
          higher.update_Q(current_state, next_state);

          if ( i%rounds==0 &&
               ((T+ratio_T) > (R+ratio_R)) &&
               ((R+ratio_R) > (P+ratio_P)) &&
               ((P+ratio_P) > (S+ratio_S)) &&
               (2*(R+ratio_R) > ((T+ratio_T) + (S+ratio_S)))
          ){
            R = R + ratio_R;
            T = T + ratio_T;
            S = S + ratio_S;
            P = P + ratio_P;
            higher.setPayoff_matrix(R, S, T, P);
            lower.setPayoff_matrix(R, S, T, P);
            System.out.printf("R:%.4f T:%.4f S:%.4f P:%.4f e:%d t:%d\n",
higher.getPayoff_matrix(0),
                    higher.getPayoff_matrix(2), higher.getPayoff_matrix(1),
                    higher.getPayoff_matrix(3), i, t);
          }

          current_state = next_state;

          update_states_results(episodes_before+i-1, current_state, t);
          update_payoff_results(episodes_before+i-1, lower.getPayoff_matrix(current_state),
higher.getPayoff_matrix(current_state), t);
        }
      }

      statistics(trials, (episodes_before+episodes_after));
      printToFile_statistics(trials, (episodes_before+episodes_after));

    }
}
```

Appendix B

Class Player.java

```
package com.example.QlearningModel;

/**
 * Created by jeannettechahwan on 20/01/15.
 */

public class Player {

   private double learning_rate;
   private double epsilon;
   private double discount;
   private double[][] Q_table;
   private double[] payoff_matrix;
   private int current_action;

   public Player(){
      this.Q_table = new double[4][2];
      this.payoff_matrix = new double[4];
   }

   public void setDiscount(double value){
      this.discount = value;
   }

   public double getDiscount(){
      return this.discount;
   }

   public void setEpsilon(double value){
      this.epsilon = value;
   }

   public double getEpsilon(){
      return this.epsilon;
   }

   public void setLearning_rate(double value){
      this.learning_rate = value;
   }

   public double getLearning_rate(){
      return this.learning_rate;
   }

   public void setQ_table(int row, int column, double value){
      this.Q_table[row][column] = value;
   }

   public double getQ_table(int row, int column){
      return this.Q_table[row][column];
   }
```

```
  public void setPayoff_matrix(double value1, double value2, double value3, double value4){
    payoff_matrix[0] = value1;
    payoff_matrix[1] = value2;
    payoff_matrix[2] = value3;
    payoff_matrix[3] = value4;
  }

  public double getPayoff_matrix(int row){
    return this.payoff_matrix[row];
  }

  public void setCurrent_action(int action){
    this.current_action = action;
  }

  public int getCurrent_action(){
    return this.current_action ;
  }

  public int choose_action(int state){
    int action;
    double random = Math.random();

    if(random<epsilon){   //explore
      action = (int) (Math.random() * 2);   // values: 0(cooperate) or 1(defect)
    }

    else {   //exploit
      // find best known action
      if (this.getQ_table(state, 0) > this.getQ_table(state, 1))
        action = 0;
      else if(this.getQ_table(state, 0) < this.getQ_table(state, 1))
        action = 1;
      else
        action = (int) (Math.random() * 2);   // values: 0(cooperate) or 1(defect)
    }

    return  action;
  }

  public void update_Q(int current_state, int next_state){
    double max, Q;

    // find maximum value from Q table
    if(this.getQ_table(next_state,0) >= this.getQ_table(next_state,1))
      max = this.getQ_table(next_state, 0);
    else
      max = this.getQ_table(next_state, 1);

    // Calculate Q and set value in player's Q table
    Q = ((1 - this.learning_rate) * this.getQ_table(current_state, this.getCurrent_action())) +
(this.learning_rate * (this.getPayoff_matrix(next_state) + (this.getDiscount() * max)));
    this.setQ_table(current_state, this.getCurrent_action(), Q);
  }

}
```