TÉCNICO LISBOA

Overcoming automatic response tendencies: behavioral findings and computational model-based analysis

Tiago Alexandre Delgado Henriques

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisors: Prof. Tiago Vaz Maia

Prof. Patrícia Margarida Piedade Figueiredo

Examination Committee

Chairperson: Prof. Ana Luísa Nobre Fred

Supervisor: Prof. Tiago Vaz Maia

Member of the Committee: Prof. João Miguel Raposo Sanches

December 2015

Acknowledgements

Before starting to address this thesis' theoretical sections, I would like to thank to several people who inspired me and helped me getting where I am today. To anyone I forgot to mention a big "bem-haja".

Moreover, I want to thank to my special family, not only for their support throughout this journey, but also for believing in my work and helping me be the person I am today.

First of all, I would like to thank Professor Tiago Vaz Maia for the long and enriching meetings as well as for the opportunity of working in his lab and meeting such incredible people, that not only helped me through their knowledge but also with their support. A special thank you to postdoctoral fellow Lena Ernst who became a mentor and also a close friend; to Vasco Conceição for all his advice, remarks and complements which were essential on helping me to grow as an individual, as well as developing a keen and scientific mind. Likewise I have to mention the discussions I had with Ângelo Dias, Catarina Farinha, Ana Portelo, Inês Santa Ana e Rita Belo and the advice given by them.

In addition, I have to express my gratitude to all experimental participants, because without them this thesis would not exist. Once again, thank you!

Furthermore, I would like to thank as well to all my dearest friends who were always there throughout my academic journey and made it unique and special in different ways: Ana Sousa, Inês Vicente, Daniel Almeida, Felipe Henriques, Gil Luís, Helena Forte, Tânia Agostinho, Gonçalo Gomes, Duarte Mendes de Almeida, Pedro Parreira, Daniela Pinheiro and others. To Élson Tomás a special thank you for his friendship, all his support and every single one of his remarks and advice.

I also want to express my gratitude regarding my friend, confidant and girlfriend Márcia Luzia for all her love, inspiration, energy, motivation and encouragement given all these years.

To conclude, I want to thank my grandparents who are no longer here, but who definitely watched out for me and provided me inspiration and motivation in order to finish this phase. Especially to my grandfather António, a huge thank you for the inspiration and showing me the willingness to live, even when everything is against us. To you, my life mentor, a final gesture of gratitude.

Abstract

Basic dimensions of behavior, namely approaching positive and avoiding aversive stimuli correspond to evolutionary biases that confer survival advantages. However, frequently, these automatic reactions jeopardize our social integrity. Therefore, we must be able to engage in incongruent responses (approaching negative and avoiding positive stimuli) which are effortful and depend on considerable cognitive resources.

This thesis aimed to verify whether the repetitive performance of such incongruent reactions would automatize them, *i.e.*, would lead to habit-like behavior. Ideally, such high frequency pairing would build-up a stable *stimuli-incongruent reaction* association. To test this hypothesis, healthy participants were asked to perform a computerized Approach-Avoidance task, during a 5 consecutive-day training period, and novel computational models were developed to fit participants' reaction times.

The model that best fitted the data was selected through Bayesian approaches and it was proven to be significantly better than a model assuming that the average value of the reaction times did not change during training. The selected model presented six free parameters which captured processes involved in habit learning and cognitive control, as well as Pavlovian biases.

Further analyses showed participants to perceive negative pictures significantly less aversively after the training period, which might be a consequence of habit formation, as transduced by the significant decrease of reaction times for the *approach negative* condition.

In the future, similar training protocols might be an add-on therapy in patients with obsessive-compulsive disorder, since current exposure and response prevention therapies are uncomfortable, time-consuming and associated with a high relapse-rate.

Keywords

Approach-Avoidance task; Behavioral modeling; Habit formation; Incongruent reactions.

Resumo

As bases do comportamento, nomeadamente aproximar estímulos positivos e evitar estímulos negativos, correspondem a tendências que conferem vantagens evolutivas. Contudo, estas reações automáticas, frequentemente, põem em causa a nossa integridade social. Deste modo, devemos ter a capacidade de apresentar respostas contraditórias (aproximar estímulos negativos e evitar estímulos positivos) que exigem um grande esforço cognitivo.

O objetivo desta tese é verificar se a prática recorrente de tais respostas torna-as automáticas, conduzindo em última instância à habituação. Idealmente, esta repetição conduz ao estabelecimento de associações estáveis de *estímulo-reação* contraditória. Para provar esta hipótese, pedimos a participantes saudáveis que executassem uma *Approach-Avoidance task* computorizada, ao longo de 5 dias consecutivos de treino, e desenvolvemos novos modelos computacionais que se ajustaram aos tempos de reação adquiridos.

O modelo que melhor se ajustou aos dados foi selecionado através de abordagens Bayesianas e demonstrou ser significativamente melhor que um modelo que assumia tempos de reação constantes ao longo do treino. O modelo selecionado, caraterizado por seis parâmetros, descreve processos relacionados com *habit learning* e controlo cognitivo, assim como princípios pavlovianos.

Análises posteriores demonstraram que os participantes consideraram as imagens negativas apresentadas significativamente menos aversivas após o período de treino, possível consequência da habituação, traduzida pelo decréscimo significativo dos tempos de reação associados à condição *approach negative.*

Futuramente, treinos semelhantes poderão ser protocolados e utilizados como terapia para pessoas que apresentem transtornos obsessivo-compulsivos, uma vez que a atual terapia de exposição e prevenção de resposta é incómoda, demorada e está associada a uma alta taxa de reincidência.

Palavras-chave

Approach-Avoidance task; Habituação; Modelação comportamental; Reações contraditórias.

Contents

List of Figures

4.2 Representation of the output provided by function *histDist*, on the behavioral data of the 8th subject of the positive group. The blue line is an interpolation performed to the histogram and the red line in (a) represents the fitted Normal distribution whose parameters (μ and σ) were estimated from the RTs and (b) represents the fitted Log-Normal distribution whose parameters (μ and σ) were estimated from the RTs.

4.3 Representation of the output provided by function *histDist* on the behavioral data of the 8th subject of the negative group. The blue line is an interpolation performed to the histogram and the red line represents the fitted Normal distribution whose parameters (μ and σ) were estimated from the RTs in (a) the first session, (b) the second session, (c) the third session, (d) the fourth session and (e) the fifth session.

4.4 Representation of the output provided by function *histDist* on the behavioral data of the 8th subject of the negative group. The blue line is an interpolation performed to the histogram and the red line represents the fitted Log-Normal distribution whose parameters (μ and σ) were estimated from the RTs in (a) the first session, (b) the second session, (c) the third session, (d) the fourth session and (e) the fifth session.

4.5 Graphical results of the analysis of the arrow version of the AAT. (a) Depicts the mean RTs in ms, for the different actions within each group, and (b) depicts the ratio between the reaction times associated with avoidance and approach reactions.

4.6 Results from the assessment version of the AAT that depict the reaction biases before and after the training of the positive group, towards negative and neutral pictures they did not train, and for the negative group, towards the positive and neutral pictures they did not train (left side), and reaction biases before and after the training of positive group, towards the conditions they trained, and for the negative group, towards the conditions they trained (right side).

4.7 Results from the assessment version of the AAT. The differences in reaction biases before and after training of the positive group, towards negative and neutral pictures they did not train, and for the negative group, towards the positive and neutral pictures they did not train (left side), and the differences in reaction biases before and after training of positive group, towards the conditions they trained, and for the negative group, towards the conditions they trained (right side).

4.8 Results from the assessment version of the AAT. The differences in reaction biases before and after training of the positive group, towards negative and neutral pictures they did not train, and for the negative group, towards the positive and neutral pictures they did not train (the three initial bars) and the differences in reaction biases before and after training of positive group, towards the conditions they trained, and for the negative group, towards the conditions they trained, including results obtained for generalization, respectively (right side).

4.9 Results of the linear fit performed to each condition using the mean of the reaction times between pictures of that condition and between the subjects that trained that condition.

4.10 Results of the exponential fit performed to each condition using the mean of the reaction times between pictures of that condition and between the subjects that trained that condition.

4.11 Results of the power law fit performed to each condition using the mean of the reaction times between pictures of that condition and between the subjects that trained that condition.

4.12 Results of the exponential model's fit performed to the behavioral data at a group level.

4.13 Results of the power law model's fit performed to the behavioral data at a group level.

4.14 Evolution of the ratings along 5 days of training of the pictures trained in the 4 different conditions.

4.15 Evolution of the individual ratings along 5 days of training. (a) Depicts the evolution of the 7th participant of the negative group, (b) depicts the evolution of the 14th participant of the negative group, (c) depicts the evolution of the 3rd participant of the positive group and (d) depicts this evolution of the 4th participant of the positive group. The different less marked dashed lines correspond to the evolution of the trained pictures.

4.16 Ratings provided by the participants before and after the training of the positive group, towards negative and neutral pictures they did not train, and for the negative group, towards the positive and neutral pictures they did not train (left side), and ratings before and after the training of positive group, towards the conditions they trained, and for the negative group, towards the conditions they trained, including the generalization results, respectively (right side).

4.17 Results from the practical test, which have depicted the mean RTs in seconds (s) that participants took to sit down on a specific category of picture for each group.

4.18 Results of BMC between the models described in table 3.2 obtained using the BIC, for the population. (Top Left) Model Attribution at subject level, the first eighteen subjects belong to negative group while the last eighteen to the positive group. (Top Right) Estimated model frequencies. (Left and Right Bottom) Approximated Exceedance Probabilities (EPs) and Protected EPs of the eight models.

4.19 Histogram of the Pavlovian parameter estimated by the computational model selected through BMS. The vertical dashed grey line identifies the value zero.

4.20 RTs predicted by the computational model for 4th subject of the positive group. The red line represents the RTs acquired and the blue line the RTs predicted by the computational model.

4.21 Results of BMC between the models which received as input transformed RTs (via moving average filtering) described in table 3.2 obtained using the BIC, for the population. (Top Left) Model Attribution at subject level, the first eighteen subjects belong to negative group while the last eighteen to the positive group. (Top Right) Estimated model frequencies. (Left and Right Bottom) Approximated Exceedance Probabilities (EPs) and Protected EPs of the eight models.

4.22 Histogram of the Pavlovian parameter estimated by the computational model selected through BMS, considering the application of the moving average method on the raw data. The vertical dashed grey line identifies the value zero.

A1.1 Structure of the excel file Test.xlsx.

A1.2 Dialogue box displayed at the beginning of the task (a) to select the excel input file and (b) to fill in the blank boxes where the experimenter must insert the subject's number, the number of session and number associated to the routine that will be performed.

A2.1 Results from the assessment version of the AAT that depict the reaction biases before and after the training of the positive group, towards negative and neutral pictures they did not train, and for the negative group, towards the positive and neutral pictures they did not train (left side), and reaction biases before and after the training of positive group, towards the conditions they trained, and for the negative group, towards the conditions they trained (right side).

A2.2 Results from the assessment version of the AAT. The differences in reaction biases before and after training of the positive group, towards negative and neutral pictures they did not train, and for the negative group, towards the positive and neutral pictures they did not train (left side), and the differences in reaction biases before and after training of positive group, towards the conditions they trained, and for the negative group, towards the conditions they trained (right side).

A2.3 Results from the assessment version of the AAT. The differences in reaction biases before and after training of the positive group, towards negative and neutral pictures they did not train, and for the negative group, towards the positive and neutral pictures they did not train (the three initial bars) and the differences in reaction biases before and after training of positive group, towards the conditions they trained, and for the negative group, towards the conditions they trained, including results obtained for generalization, respectively (right side).

A2.4 Results of the linear fit performed to each condition using the mean of the reaction times between pictures of that condition and between the subjects that trained that condition.

A2.5 Results of the exponential fit performed to each condition using the mean of the reaction times between pictures of that condition and between the subjects that trained that condition.

A2.6 Results of the power law fit performed to each condition using the mean of the reaction times between pictures of that condition and between the subjects that trained that condition.

A4.1 Individual RTs acquired by the arrow version of the AAT.

A4.2 Individual RTs acquired by the arrow version of the AAT for the negative group.

A4.3 Individual RTs acquired by the arrow version of the AAT for the positive group.

A5.1 Results of BMC between the first four models which received as input transformed RTs (via moving average filtering) described in table 3.2 obtained using the BIC, for the population. (Top Left) Model Attribution at subject level, the first eighteen subjects belong to negative group, while the last eighteen to the positive group. (Top Right) Estimated model frequencies. (Left and Right Bottom) Approximated Exceedance Probabilities (EPs) and Protected EPs of the eight models.

List of Tables

List of Abbreviations

AAT Approach Avoidance Task

ACC Anterior Cingulate Cortex

AIC Akaike Information Criterion

BAS Behavioral Approach System

BG Basal Ganglia

BIC Bayesian Information Criterion

BIS Behavioral Inhibition System

BMC Bayesian Model Comparison

BMS Bayesian Model Selection

BOR Bayesian Omnibus Risk

CI Confidence Interval

CNS Central Nervous System

DA Dopamine

DLPFC Dorsolateral Pre-Frontal Cortex

EPs Exceedance probabilities

FFFS Fight/ Flight/ Freeze System

FMRI Functional Magnetic Resonance Imaging

GP Globus Pallidus

L-BFGS Limited memory Broyden–Fletcher–Goldfarb–Shanno

LLH Log-likelihood

ME Model Evidence

MS Model Evidence

OCD Obsessive-Compulsive Disorder

OFC Orbitofrontal cortex

OLS Ordinary Least Squares

PEPs Protected Exceedance probabilities

PFC Pre-Frontal Cortex

RIM Reflective Impulse Model

RL Reinforcement Learning

RTs Reaction Times

RW Rescorla-Wagner

SD Standard Deviation

SN Substantia Nigra

STN Subthalamic Nucleous

S-R Stimulus-Response

SRC Stimulus Response Compatibility

S-R-O Stimulus-Response-Outcome

SSE Residual sum of squares

VLPFC Ventrolateral Pre-Frontal Cortex

WM Working Memory

1

Introduction

Contents

1.1. Motivation

"*A esperança de ter um impacto direto e significativo na melhoria das vidas de milhões de pessoas que sofrem todos os dias com os problemas que estudo*".

The hope to have a direct and significantly impact on the lives of millions of people that suffer every day from the problems that I study

This quote from Professor Tiago Maia summarizes perfectly my enthusiasm to embrace this thesis.

Even though we know this project will not have such an immediate impact, we believe this is the beginning of something that might actually change the daily life of many in a future not so far away.

The current thesis is part of a bigger project led by Professor Tiago Maia and the postdoctoral fellow Lena Ernst, whose main objective is to present a daily add-on therapy based on the performed Approach-Avoidance Task (AAT) training. This therapy is supposed to simplify and fortify the standard Exposure and Response Prevention (ExRP) treatment, by facilitating approach reactions of patients with obsessive-compulsive disorders (OCD) to feared stimuli. The ExRP is the psychotherapeutic treatment of choice for OCD patients due to its efficacy and high empirical support. However, besides of the high relapse rate, it is a substantially discomforting and time-consuming treatment [1] [2] [3]. To find ways to make this treatment less aversive firstly, it is relevant to identify the mechanisms that are affected by this disorder. OCD patients have predisposition toward excessive stereotyped behavior in order to avoid adverse consequences [4], therefore it is broadly accepted that avoidance rather than appetitive compulsions are characteristic of OCD patients [5]. This in turn led us to assume that some OCD symptoms might be seen as enhanced avoidance tendencies [6] [7] [8].

Before going further into details regarding OCD patients, we first have to understand what is happening with healthy people, when asked to perform basic, yet vital, behaviors such as approaching positive and avoiding negative stimuli. These automatically triggered tendencies correspond to evolutionary biases that generally confer a survival advantage. Notwithstanding, taking into consideration that we are part of a civilized society, we have to respect social conventions and behave appropriately. Therefore, when faced with situations that require responses that go against our most primitive tendencies, human beings must be able to engage in so-called controlled responses [9]. In fact, these forms of behavior are as important as the automatic ones.

Although, humans are able to regulate certain responses, these responses are effortful and very demanding concerning willpower and attention [10]. In fact, the continuous allocation of our cognitive and motivational resources to overcome automatic response tendencies might lead to the incapability to sustain controlled responses, due to exhaustion [11]. Therefore, finding ways to make beneficial controlled responses themselves more automatic is mandatory, even when those responses go on opposite directions than our pre-determined biases.

Bearing in mind that the brain is one of the most complex structures known, the identification of a correct model to understand the psychological and cognitive processes used by humans when performing an incongruent action (approaching the negative or avoiding the positive) will be of utmost importance, especially because this might help to disentangle the computations performed by the brain at a higher level. Only after this, we can start to understand how pathophysiological processes alter

these computations, leading to patients' symptoms. Therefore, it is obvious that standard approaches will not be enough to handle this kind of challenges [12].

So, computational models emerge at the forefront to overcome this challenge not only because of the outstanding work that has been developed in this area, specifically in Reinforcement Learning [12] [13] [14], but also because there is strong evidence of a correlation between the variables computed by these models and the neuronal function [15] [16].

Therefore, we expect that the application of novel computational methods, combined with the new developments which have been made regarding parameter estimation and model comparison, will be essential to get a better understanding of the mechanisms involved in the aforementioned approach and avoidance processes. Still, it is very important to be cautious when applying these new methods, since their misuse might lead to highly unreliable conclusions.

1.2. Objectives

This thesis' main goal is to understand the mechanisms of acquiring new habits that contradict automatic response tendencies, by disentangling and quantifying the different cognitive processes involved with computational learning models.

Therefore through the development of a novel versions of the AAT, the application of a practical test and the acquisition of the individual ratings, we will assess behavioral effects of training participants to perform incongruent actions, such as approaching negative stimuli and avoiding positive ones.

As a training effect we expect changes in the participants' reaction times (RTs). Besides, these changes might influence the way participants feel about certain stimuli, so we will also assess the ratings of pictures used. Moreover, in order to capture subjects' behavior during the training task, and present as parameters internal variables correlated with different psychological processes, we will apply novel computational models to the acquired data.

1.3. Thesis Outline

This thesis is composed by five main chapters.

The second chapter provides biological, psychological and physiological information which is the key to understand the mechanisms we aim to investigate. Moreover, there we focus on the habit learning component, more specifically on how stimulus-response associations are affected by repetition, since this is one of the matters we aim to study.

In the third chapter, we provide more background regarding the explanation of the methods used for this study. We start to discuss the AAT, its implementation and remarks regarding the different versions. Then, we will point out some pre-processing processes that were useful in order to reduce the noise in the data. Next we introduce the computational models used to obtain computational biomarkers for the different psychological and cognitive processes, and the set of *a priori* hypotheses to be tested. We terminate this chapter by explaining the methods used to estimate the parameters of these models and to perform model comparison.

The fourth chapter will present and discuss the results obtained from the exploratory data analysis and the results from the model-free and model-based analyses. Here, we make use of mixed models in order

to test the hypotheses we presented for the datasets obtained with the different versions of the AAT, and we analyze the results obtained from the practical test. This chapter finalizes with the results from the model selection procedure applied to the different computational models and with some parametric tests.

The final chapter is composed by remarks regarding the work developed and by possible future developments and directions that might follow this project.

1.4. State of the art

During this thesis, we will combine behavioral experimentation with neuro-computational models that will try to (partially) explain some of the cognitive and psychological processes involved in approach and avoidance behaviors. Next, we present a brief description of the most recent findings in these two areas.

1.4.1. The Approach-Avoidance Task (AAT)

Behaviorally, this thesis focuses on the interplay between impulses and their inhibition through self-control [11]. By using the AAT we aim to disentangle this complex behavior to identify the cognitive subcomponents as was done in dieting [17] and smoking cessation [18] studies with regard to working memory (WM) capacity and its limits.

The AAT is an extensively used implicit task since it directly assesses the behavioral component of approach-avoidance impulses and also allows to evaluate the deliberative regulation of these impulses [19].

Since the first experimental investigation on arm movements and their relation to affective processes, researchers have used a variety of experimental tasks simulating approach-avoidance behavior to investigate these processes. In 1960, Solarz described a plausible influence of the compatibility of the stimulus valence and automatic approach-avoidance tendencies on RTs. In this study, participants had to move a hand lever to approach or avoid word cards categorized as positive and negative stimuli. The results revealed that participants were faster to initiate compatible movements (approach positive and avoid negative words) than incompatible ones [20].

Then, in 1999, Chen and Bargh and, later on, Rinck and Becker modified the AAT into its classical, computerized version, allowing to investigate automatic approach-avoidance tendencies and their regulation by cognitive control in healthy subjects [21].

This version has also proven useful to understand certain aspects of psychopathology, such as the pathologically enhanced approach tendencies for alcohol-related stimuli in alcohol dependence [22] [23] and for heroin-related stimuli in people addicted to heroin [24], and the pathologically enhanced avoidance tendencies for phobia-relevant stimuli in different phobias [25].

For this reason, a training version of the AAT was developed to help patients with alcohol dependence overcome their tendency to approach alcohol-related stimuli. The results were promising, as the patients who underwent this training showed a tendency for reduced relapse probability after one year [26].

Besides this, Najmi *et al.* (2010) used the joystick AAT to investigate the responses to threatening stimuli of two distinct groups, the first with participants high in contamination-related obsessive-compulsive symptoms (HC) and the second group with participants low in contamination-related

obsessive-compulsive symptoms (LC). They made participants approach and avoid both contamination-related and neutral pictures, and the results showed that participants of the HC group were significantly faster to approach neutral pictures than to approach contamination-related pictures while the other participants did not present a significant difference in approaching neutral and contamination-related pictures. More interestingly, they also presented a significant correlation between the slowliness of avoidance when pulling contamination-related stimuli and self-report contamination-related obsessive-compulsive symptoms. These findings were interpreted to provide evidence that a biased behavior response was only found when participants of the HC group tried to inhibit the automatic response of avoiding threatening stimuli. Therefore the results of this study validated the use of the AAT to quantify the inhibited and uninhibited prepotent avoidance reactions to threatening stimuli in participants with contamination-related obsessive-compulsive symptoms [27].

Following the results of the last study, Amir *et al.* (2013) used the joystick version of the AAT to investigate the impact of manipulating automatic response tendencies. To do so, they conducted a study where healthy participants were trained to approach contamination-related pictures. Then, participants were asked to approach and avoid neutral and contamination-related pictures. The results showed that the manipulated participants presented facilitated approach responses and reduced avoidance tendencies towards contamination-related stimuli when compared to a not-trained control group. These findings have serious clinical implications since they showed evidence for a modification of automatic action tendencies. These results indicate a decrease of the automatic avoidance tendency for contamination-related stimuli, in individuals with contamination-related symptoms [28], that could also be interpreted as an increase in the cognitive effort employed.

Considering the aforementioned studies and their respective results, it is also important to mention that attempts to unravel what was underlying the efficacy of the referred training have already been made. Sharbanee *et al.* (2014) investigated if the training's effect was mediated by a change in action tendency or a change in selective attention and whether it was moderated by individual differences in WM capacity. The authors assessed whether the impact of the AAT training on alcohol consumption was mediated by its impact on alcohol action tendency, rather than by its impact on selective attention to alcohol, using a recently developed Selective-Attention/ Action Tendency Task (SA/ATT). Moreover, they also evaluated WMC as a potential moderator of the training efficacy.

Although, there was a significant indirect effect of training on alcohol consumption mediated by a change in action tendency, no effect was mediated by a change in selective attention. There was inconsistent evidence of WMC moderating training efficacy, with moderation found only for the effect of the approach-alcohol training on the AAT but not on the SA/ATT. Thus, the results suggested that approach/avoidance training affects alcohol consumption by changing the underlying action tendency [29].

At the same time, Radke *et al.* (2014) used the same task to investigate action tendencies in patients with depression, since individuals with depression present approach deficits including anhedonia, reduced energy and social withdrawal.

In this study, it was used an explicit zooming version of the AAT, where participants had to react to emotional faces by either pushing a joystick away from them or pulling it towards them, simulating the avoidance and approach reactions respectively.

The results indicated that behavioral adjustments to different emotional expressions, gaze directions or motivational demands were lacking in depression. Remarkably, it allowed to distinguish depressed patients not only from healthy individuals, but also from other clinical populations that demonstrate aberrant approach–avoidance tendencies, *e.g.*, from patients with social anxiety or psychopathy [30].

However, besides the above mentioned study that tried to investigate the influence of the WM capacity in the training AAT, neither the cognitive processes nor other neuronal substrates involved in the same routine have been investigated thus far.

In fact, using the classical version of the AAT, it has been shown that incompatible, controlled responses compared to compatible, automatic reactions elicit stronger activation in a specific brain area, the dorsolateral prefrontal cortex (DLPFC) [31]. Even though this and similar studies are interesting and important, they do not offer practical insights into the prevention of impulsive action.

However, by adding the substantial body of work highlighting the usefulness of computational reinforcement learning models in characterizing behavior and brain-behavior relationships to the aforementioned findings, we expect that it will be possible to create the necessary framework for the characterization of relevant aspects of the cognitive processes involved in the AAT [12] [32].

1.4.2. Computer Science: Learning Models

"*In the last 15 years, there has been a flourishing of research into the neural basis of reinforcement learning, drawing together insights and findings from psychology, computer science, and neuroscience*." [14].

This proliferation of research exploring the neural and psychological mechanisms of reinforcement learning has focused mainly on the decision making process used by animals and humans when selecting actions in the face of rewards and punishments [32] [33]. Thus, before going into more complex value-based decision making in humans, it is important to mention two paradigms by which these processes are ruled:

- Classical conditioning: the subject learns an association where the outcome is independent on the action. Therefore it establishes connections between stimuli.
- Instrumental conditioning, when contrarily to the first, the subject learns that the outcome depends on the action, thus it establishes connections between stimuli and actions [32].

It was precisely based on these paradigms and on several observations during the last decade that a good parallel between neurobiological processes in the brain and the computational Reinforcement Learning[1] (RL) models was established.

In fact an extraordinary finding was the discovery of the relationship between dopamine and prediction errors. Specifically, it was shown that phasic responses of dopaminergic neurons in the midbrain code

[1] Reinforcement Learning (RL) is a branch of artificial intelligence. Moreover, it is a type of learning that is concerned about how an agent learns to optimize decisions through the feedback given by the surrounding environment, in order to maximize rewards [34].

reinforcement prediction errors [15] [35]. Another important finding was that the plasticity of corticostriatal synapses is weighted by dopamine input from midbrain dopamine neurons [36].

This provided a normative framework to understand animal and human conditioned behavior, a process based on two steps: reward estimation and action choice. Later on, in 1972, in order to translate mathematically the idea that a subject learns through prediction errors, *i.e.*, through the difference between the actual outcome and the expected one, Rescorla Wagner developed the first RL model through equation 1.1.

$$V_{new} = V_{old} + \alpha(r - V_{old}) \tag{1.1}$$

Here the r represents the reward on a given trial, V the associative value of the stimulus and α is the learning rate that controls the degree of update.

Since then, much has been done regarding the development of this model, since its original form was not able to predict more complex behaviors (for further details see [14]).

These developments combined with the findings above mentioned gave rise to the basal ganglia Go/NoGo neuro-computational model which explains how temporal difference methods are implemented in the basal ganglia circuitry and how they influence action selection [32] [33] [35].

Although all of these findings are really outstanding, very little was found regarding the applicability of concepts of reinforcement learning models to the scope of this thesis. Nonetheless, it is worth to mention the work of Palminteri *et al.* (2011) who employed a novel paradigm to demonstrate that positive feedback can improve motor skill learning in humans. In fact, the author found that healthy participants progressively got faster in executing sequences of key presses that were followed by positive feedback.

Once again it is important to state that this was the first study to investigate the influence of reinforcements on motor skill learning and to give empirical evidence that providing reinforcements helps people acquiring motor skills. Besides, the authors also developed a computational framework in which the prediction of a positive feedback facilitated the execution of the desired action [37].

Nevertheless, the development of computational models that capture psychological and cognitive processes which Palminteri's work did not capture, such as the cognitive effort employed in task solving, is of utmost importance.

Besides, several studies, which reported action tendency modification through the use of the AAT [23] [24] [26] [28], support this requirement. These studies also lead us to hypothesize that people are learning to modify their automatic responses through habit formation (cf. sub-section 2.4.1) , which results from the repetition of a specific instructed action in a specific context, and that novel computational models will be essential to study this modification.

2

Background

Contents

In this chapter we briefly introduce the concepts of approach and avoidance along with physiological and anatomical findings concerning three brain areas that play major roles in these behaviors.

Then, an overall review of the theories that correlate these forms of behavior and the physiological and anatomical findings is given.

Next, we provide insights regarding regulated processes and related neuronal models, as well as instrumental learning.

Finally some details and regards of the use of the AAT are provided.

2.1. Basic principles of approach and avoidance behavior

Approach and avoidance reactions are seen as essential behavioral principles, as they are considered a basic dimension of the maintenance of human homeostasis[2] [9]. This is assumed to be due to the fact that we are constantly exposed to a panoply of stimuli whose nature may derive from the surrounding environment or from stimuli caused by the organism itself, which might result in changes of the human body reactions. Inherently, this evokes feelings which trigger affectively enriched feedbacks that allow the detection of these changes and induce appropriate reactions.

Therefore, these behavioral reactions, more specifically approaching positive and avoiding negative stimuli, are common to many living beings and most of the times automatically triggered due to evolutionary biases that confer a survival advantage.

Human beings, however, have a much broader spectrum of behavioral abilities, which allow them to regulate these tendencies and engage in controlled responses that may be against their primary instincts but are advantageous for the achievement of long term-goals. The practical repercussions of this regulation ability are huge since we are daily confronted with situations in which the automatic response tendencies cause hedonic fulfillment in the short-term but harm in the long-term [19].

In the following sections, we provide an overview of the psychological constructs, present investigations on neuronal correlates and then a detailed look at the underlying mechanisms.

2.1.1. Approach the positive and avoid the negative

The basic relations between stimulus categorization and behavior has been extensively studied: Significant evidence in experimental psychology confirms that, in the presence of positive stimuli, approach behavior is facilitated, while facing negative stimuli facilitates avoidance behavior [21]. Moreover this relation is bidirectional, since the categorization of positive stimuli as “positive” is facilitated when approach related behavior is performed (*e.g.* flexion of the arm) and in the other case, the identification of a negative stimulus is facilitated by avoidance related behavior (e.g. stretch the arm) [38].

Bearing these things in mind, it is generally agreed that positive approach and negative avoidance assignments are considered congruent pairings while positive-avoidance and negative-approach assignments are interpreted as incongruent pairings; a better performance is associated with congruent stimuli response pairings. Even though, there is a huge diversity of emotional and motivational theories

[2]This term refers to a process that maintains the stability of the human body's internal environment in response to changes in external conditions [39].

describing aspects of great relevance to understand this phenomenon, it will not be scrutinized since it is not the scope of this thesis. Nevertheless, there is evidence from experimental psychology that the stimulus categorization either as positive or negative occurs automatically and the functional value of such automatic attitude activation is preponderant in the reaction to a complex environment. Besides there is evidence that the evaluation of stimuli and related approach-avoidance behavior might be different across people due to inter-individual personality differences [21].

Nonetheless human behavior is assumed to be a function of both psychological and environment constructs (e.g. the needs and features of an object), thus the valence assigned to a stimulus depends strongly on the two factors mentioned. These in turn determine the general direction of behavior by evoking motivational forces (this term refers to the incentive to produce an action, based on internal needs, cognitive motives and external stimuli). Indeed, Lang *et al.* (1990) stated that emotions are action dispositions that prepare our organisms to respond quickly and appropriately [40] [41].

In fact, findings that associated the vertical nodding of the head with positive affect and approach motivation and the horizontally shacking of the head with negative affect and avoidance motivation were already reported [21].

Besides that, there are also two muscles of specific interest for approach-avoidance behavior: the flexor muscle of the arm, responsible for bending it and the tensor responsible for stretching it. Actually, the flexion and extension of the biceps are assumed to be distinctively linked to approach and avoidance behavior, due to a long life higher order Pavlovian conditioning process, because the repetition of these specific movements during an individual's life is intrinsically connected with approaching desirable goods and avoiding undesirable goods, respectively [38] [42].

Several further studies have also suggested that the regulation of the distance between the subject and the stimulus is more relevant as the muscle activation. Indeed, Markman and Brendl (2005) showed that the reference point to which the stimulus is being approached or avoided plays a crucial role to interpret the relation between the stimulus valence and behavior. This means that only when the subject's body was the reference point, the movement of arm flexion and extension was associated to approach and avoidance behaviors, respectively [43].

2.2. Behavioral, anatomical and physiological findings

The definition of approaching positive stimuli and avoiding negative ones as vital behaviors was based on the fact that there are specialized nervous systems that process them. Neurophysiological studies contributed to understand human defensive reflexes; these findings further supported the assumption that stimulus valence is automatically processed: Reflexes are automatic reactions that are not willingly steered [21].

Following this, it was concluded that defensive reflexes in humans might rely on the same neuronal structures which are associated to the fear circuit, specifically on the amygdala [44]. This last assumption was made based on the findings of several studies in animals that allowed to obtain information not only on brain structures related to fear and avoidance, but also on brain structures related to approach motivation.

Further investigations using lesion procedures, drug administration and brain imaging techniques showed that the release of dopamine in the nucleus accumbens/ ventral striatum (segments of the Basal Ganglia) was an important factor to explain such incentive motivation causing approach behavior.

Another part of the brain linked to regulatory control abilities in humans is the Pre-Frontal Cortex (PFC) [45]. In fact, this finding was provided by Damasio *et al.* (1994) after they correlated cases of PFC damage with impairments in regulatory control, through the recreation of the bizarre accident that Phineas P. Gage suffered from (for further details see [46]). Further electrophysiological studies showed evidence for specific approach-avoidance systems regarding hemispheric asymmetry.

2.2.1. Amygdala

The amygdala is a complex structure located in the anterior temporal lobe formed by vastly interconnected nuclei (figure 2.1a). It is thought to be part of the limbic system and plays a role in many different cognitive functions, such as processing fearful and unpleasant stimuli, behavioral shaping of responses, emotional regulation, social cognition, reward processing and memory formation [44].

Regarding the nuclei that constitute the amygdala, it is important to mention the basolateral nuclear group (BLNG) (which contains the lateral, basal and accessory basal nuclei), that is interconnected with the cortex, from which it receives and converges inputs, more specifically from the sensory association cortex, PFC, and hippocampus. [45].

Another important region is the centromedial group, which contains the central and medial nuclei, and is considered to be one of the principal output regions of the amygdala. It connects mainly with the hypothalamus and the brain stem where the amygdala is responsible for hormonal and somato-motor aspects of behavior and for emotional states.

The connections of amygdala's internal nuclei start in the BLNG and end up in the centromedial group, being modulated by a small amount of GABAergic cells, known as intercalated islands, which are important because they set an inhibitory tone on amygdala pathways and modulate intrinsic passage of information from the BLNG to the central nucleus [45].

2.2.2. Basal Ganglia (BG)

The basal ganglia are a group of subcortical nuclei located in the upper region of the brain stem on both hemispheres, and they are composed by the Striatum, Globus Pallidus (GP), Substantia Nigra (SN) and Subthalamic Nucleus (STN) in humans (figure 2.1a). Its functions are maintained by three pathways from the striatum to the thalamus, namely "direct", "indirect" and "hyper-direct" pathways [32].

These three pathways implement a balance between inhibition and gating of particular representations in the cortex. More specifically, the direct pathway is responsible for inhibiting the output structures which in turn disinhibit the thalamus, enabling a specific action. The indirect pathway restrains movement through inhibition of a portion of the GP, which in turn disinhibits output structures that will suppress the thalamus. It was also suggested that the hyper-direct pathway is the moderator of these processes since it is responsible for providing tonic inhibition to the projections of the thalamus to the prefrontal cortex, thereby suppressing movement and preventing premature suboptimal responding [12][47] (figure 2.1b).

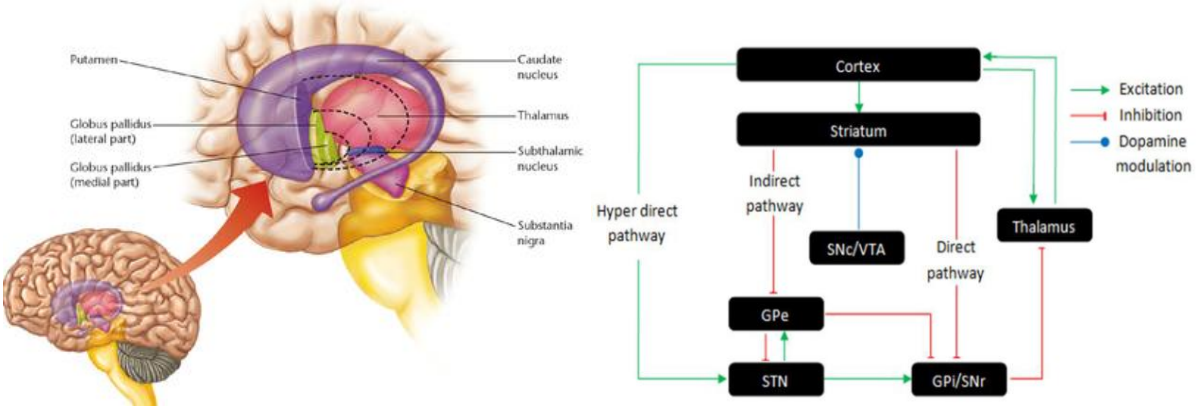


Figure 2.1: Schematics on the anatomy of the BG. (a) Location in the brain (adapted from [48]) (b) Simplified scheme of the connectivity between the structures that compose the cortico-basal ganglia-thalamo-cortical loop [49] [SNc: substantia nigra pars compacta; VTA: ventral tegmental area; GPe: globus pallidus external segment; STN: subthalamic nucleus; GPi: globus pallidus internal segment; SNr: substantia nigra pars reticulata].

Moreover, due to several different functional loops in which the BG are involved, there is strong evidence that the BG not only take part in psychological processes, but also play a crucial role in learning the relationship between sensory information and motor responses based on a trial by trial feedback [47].

2.2.3. Pre-Frontal Cortex (PFC)

The PFC represents the anterior part of the frontal lobe. This brain region has been linked to several complex cognitive functions such as planning, emotion regulation, decision making and social behavior moderation [50].

Although this brain region is divided into several parts, we consider the ones that have been consistently implicated in these functions that are the most important ones and are of interest for the current thesis: the Orbitofrontal Cortex (OFC), Dorsomedial Pre-Frontal Cortex (DMPFC), DLPFC and Ventrolateral Pre-Frontal Cortex (VLPFC) [50].

More specifically it has been proven that ventromedial areas of the PFC (the OFC and DMPFC) combined with the Anterior Cingulate Cortex (ACC) are highly connected with the amygdala. Therefore they are involved especially in the control of emotional behaviors, whereas lateral prefrontal cortical regions (that is, DLPFC and VLPFC) are mainly involved in higher executive functions such as decision making. Besides of the connection with the OFC, the DLPFC is also connected to several subcortical neural regions such as thalamus, the dorsal striatum (dorsal caudate nucleus), the hippocampus as well as primary and secondary cortical association areas [50].

2.2.4. Neuronal correlates of approach-avoidance behavior

Considering the behavior and physiological findings explained above, Davidson *et al.* (2000) proposed that the left side of the PFC is responsible for approach behavior while the other side is responsible for avoidance behavior. Further investigations highlighted that this asymmetry did not depend on the valence of the stimuli, but on distinguishing the motivational predispositions to approach or avoid a certain stimulus [51].

However, later on, it was found that in the lateral PFC there was no observable lateralization for avoidance behavior, but a left-sided lateralization for approach behavior. Besides that the medial PFC presented a pattern of a left lateralization for avoidance behavior [52].

This finding appears to contradict the aforementioned one. Nevertheless, if we focus only on the neocortical level it might be too restrictive, because the purpose of this neuronal development is to serve survival needs and the anatomy clearly demonstrates an intrinsic connection between this augmentation of the human cortex and its motivational subcortical and primitive cortical roots [53].

Meantime, another theory came up along with Davidson's theory. Based on his studies, Gray *et al.* (1994) also assumed specialized neuronal systems for approach and avoidance behavior.

Following his Reinforcement Sensitivity theory there are two emotional-motivational systems:

- The Behavioral Approach System (BAS), which deals with approach behavior to positive stimuli and security;
- The Behavioral Inhibition System (BIS), which is a system that monitors situations of mismatch between expected and current state, enabling a human being to compare and regulate certain events, *i.e.*, it helps to solve conflict situations by facilitating defensive behavior. This system contains a sub-system, the Fight/Flight/Freeze System (FFFS) that is activated by unexpected stimuli of punishment or non-reward and elicits unconditioned flight behavior and defensive aggression, *i.e.*, it is responsible for avoidance behavior [21].

Regarding the neural structures, Gray suggested the limbic system and the BA to underlie the BAS. It was suggested that the BIS is regulated by the hippocampal formation, the medial and lateral septal area and the Papez circuit[3], being the first two still influenced by the OFC. Finally, the FFFS was linked to structures associated to primary defensive reactions, namely the hypothalamus and the amygdala.

These systems are rather well-established in the broader research literature. Indeed, studies of functional Magnetic Resonance Imaging (fMRI) showed that a stronger activity in ventral striatum and OFC (structures associated with reward processing) was connected with stronger BAS, and that, in contrast, stronger hippocampus-amygdala connectivity (structures associated to punishment) was related to stronger BIS [55].

Moreover, if we take into account their interactions and how these systems differ between individuals, it is possible that all these structures form the foundations of the entire family of approach-avoidance theories [56].

2.3. Automatic and regulated processes: Impulsive and reflective systems

Considering this thesis' scope it is relevant to understand the psychology behind controlled and regulated processes, since they also play a crucial role in our daily-life, *i.e.*, if the automatic tendency we are about to execute is inadequate, then, we must be able to inhibit it and present an alternative response.

[3] The Papez circuit involves various structures of the brain. It begins and ends with the hippocampus (for further details see [54]).

Additionally, several theories in psychology and neuroscience state that human behavior is associated with the interaction between two different sets of processes: the ones that occur automatically (and in general are fast) versus the ones that are controlled and follow a plan that overcomes these automatic reactions [21].

Following the idea mentioned above, it is important to notice that several dual-system models in cognitive and social psychology consider human behavior to be supported by the interaction between these two different systems: the impulsive system and the controlled/ deliberate system [11] [18].

Furthermore most of the models developed so far relate distinct brain areas with the two systems [57], which corroborate the aforementioned Davidson's and Gray's theories. In relation to this thesis' scope, the dual system of utmost interest is the Reflective Impulse Model (RIM), where Strack and Deutsch (2004) explained how the two systems compete to determine behavior (figure 2.2).

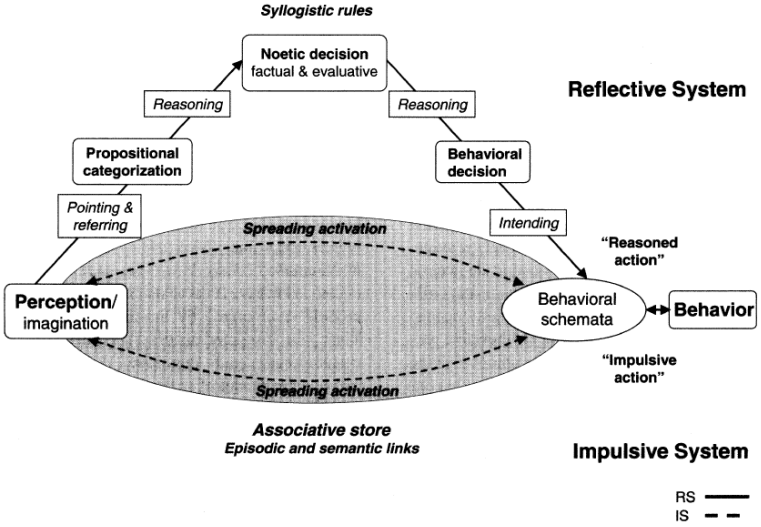


Figure 2.2: Overview of the Reflective Impulsive Model. The reflective processes are represented by full lines while the impulsive ones are represented by dash lines. Figure taken from [10].

According to this model, the impulsive system is engaged through perceptual input, therefore, it is faster and does not require much cognitive capacity. On the other hand, the reflective system provides a flexible and substantial control over decisions and actions, allowing to overcome impulsive reactions. Nonetheless, the latter is highly dependent on the allocation of control and attentional resources, meaning it will break down if these requirements are not fulfilled.

Furthermore, the output, the behavior itself, is always executed through activation of behavioral schema[4] in the motor cortex, more precisely in sensory-motor cortex clusters [11] [10].

So, basically, successful inhibition of automatic approach-avoidance tendencies requires the impulsive system's control by the reflective one, which occurs whenever several behavioral schemata (plural of schema) are activated simultaneously or when this automatic reaction has to be inhibited [31].

In healthy subjects, the reflective system controls the impulsive one in one of two ways: by deliberating taking into account the action consequences, directly imposing cognitive control on decision making, or by forcing the subject to focus on the inhibition of motor responses [21].

[4] In psychology schema describes an organized pattern of thought or behavior that organizes categories of information and the relationships among them [58].

Another relevant component regarding the control and inhibition of automatic responses is the working memory (WM) capacity. Actually, Hofman *et al.* (2008) showed that automatic attitudes towards erotic, food and alcohol-related stimuli induced a stronger influence on behavior when levels of WM capacity or the constructs related to self-control were low. This finding had already been noted by Barret *et al*. (2004) in which they concluded that individuals with a higher WMC were more successful at controlling their reactions in attention demanding circumstances [18].

Taking this into consideration it is evident that overcoming automatic reaction tendencies usually requires huge cognitive effort, since not only affective regulation and cognitive inhibition must be adequate but also the WM content [11].

2.3.1. Neuronal models

In general, it is hypothesized that controlled processes involve cortical areas, namely the PFC which has been shown to be crucial in controlling and regulating behavior [59]. In addition, it is thought that the cognitive control [5]exerted by the PFC is executed via top-down signals that influence neuronal activity in other brain areas, whilst automatic processes are supposed to be driven by subcortical regions [60].

Further studies in humans using neuroimaging methods in subjects with prefrontal lesions or in subject to whom inhibitory transcranial magnetic stimulation had been applied to (what induces temporary PFC dysfunctions), indicated that the PFC was essential for processing task relevant aspects [62][63].

However, the top-down model mentioned is unlikely to take into account all cognitive processes. In fact, the interaction between amygdala and the activation of the striatum with the PFC anatomically supports the influence of the first two structures, when emotional contents interfere with the cognitive control [45].

For this reason we concentrated on one very specific model, which describes the neural structures and respective influences on automatic approach-avoidance reactions and their regulation: the fractal triadic model. This model, suggested by Ernst and Fudge (2009), is based on the combination of three distinct, although overlapping, systems. Each of them is influenced by three sub-structures, responsible for one specific function. The inclusion of these sub-structures resulted from a variety of findings on the structures' anatomical constitution and ontogenetic development as well as on their connections with other brain areas [45].

The fractal triadic model (cf. figure 2.3) is constituted by the striatum which is mainly responsible for the approach tendencies, the amygdala responsible for the avoidance ones and the PFC moderates their activity. The striatum is characterized by its anterior part (associated to approach reactions), its posterior part (associated to avoidance reactions) and these are modulated by afferent projections from the PFC; the amygdala is characterized by the basolateral nucleus (associated to approach reactions), the central nucleus (associated to avoidance reactions) and these are also modulated by afferent projections from the PFC; the PFC is characterized by the medial OFC (associated to approach

[5] Cognitive control is considered as the capacity to assign attentional resources to steer thoughts and actions in the service of internal goals, it must be engaged to represent task-relevant information, to overcome habitual responses, to ignore irrelevant stimuli, to transform mental representations and to act in different or rapidly changing conditions [62].

processes), the lateral OFC (associated to avoidance processes) and the DLPFC is suggested to be the main structure to exert control above the other regions.

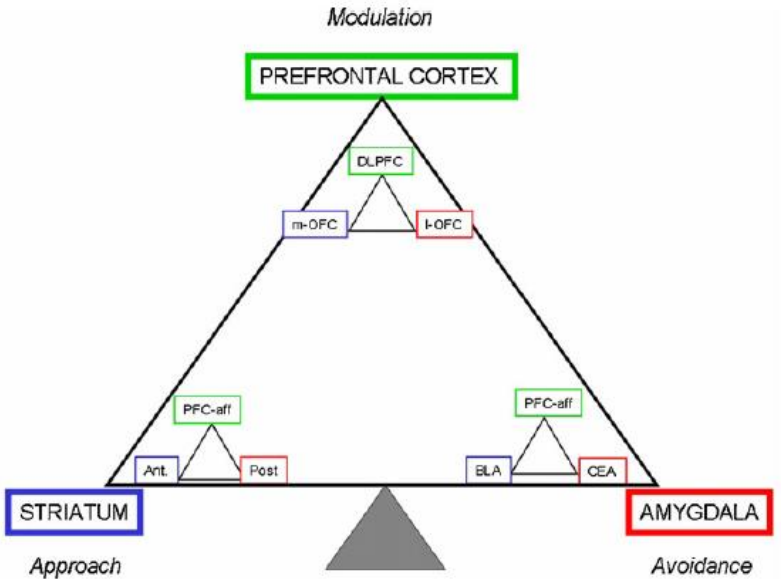


Figure 2.3: Fractal triadic model of neural systems. [DLPFC: dorsolateral prefrontal cortex; m-OFC: medial orbital frontal cortex; l-OFC: lateral orbital frontal cortex; PFC-aff: prefrontal cortical afferents; Ant: anterior striatum; Post: posterior striatum; BLA: basolateral amygdala; CEA: central amygdala]. Figure taken from [45].

Despite its many pros, Ernst *et al.* (2013) suggested an extension to this model for its application in the AAT, since it was shown that the ACC had a major contribution in conflict solving in incompatible AAT conditions [21].

2.4. Conditioning

As mentioned in sub-section 1.4.2, conditioning includes two paradigms: classical and instrumental conditioning. These will now be discussed in more detail.

The concept of classical conditioning is derived from the work of Ivan Pavlov. It occurs when a neutral or conditioned stimulus (CS) is repeatedly paired with a rewarding or unconditioned stimulus (US) given to a subject. Initially, subjects only react to the presence of the reinforcement. However after extensive training, they progressively begin to react to the CS instead of the US, as the former will predict the latter. That is, the CS will elicit a conditioned response [14].

Results from Ludvig *et al.* (2011) support the idea that classical conditioning produces stimulus-stimulus learning, rather than stimulus-response learning [14] [64]. In fact, later on results from studies by Montague, Dayan, Schultz and Sejnowski strongly sustained this idea, by establishing the association between prediction errors (PE) originated by a predictive relationship between stimuli, besides reward uncertainty, and the firing of dopaminergic neurons in the midbrain [15] [49].

Nevertheless, this form of learning could not explain how different behaviors can be performed in order to optimize the feedback received from the environment due to its limited mechanism of update [14].

From this need to explain more complex behaviors, Edward L. Thorndike formally postulated the Law of Effect, which states [65]:

"*Of several responses made to the same situation, those which are accompanied by or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur…*".

In simple terms, this law states that responses followed by an outcome would be strengthened or weakened depending on the valence of the outcome, i.e., positive outcomes would strengthen the stimulus-response (S-R) associations and negative ones would have the opposite effect. These associations are slowly learned and characterized by their rigidity since the subjects who learn this associations typically persist in performing the trained actions even when the contingencies are changed and the outcome of the trained actions is no longer of interest to them [32] [66] [67]. This reasoning was on the basis of instrumental conditioning.

However, nowadays it is known that learning does not occur just for expected rewards and not all instrumental learning is processed in terms of S-R associations (habits) since the subject's behavior might change according to the new outcome. Thus instrumental learning also includes learning which occurs through stimulus-response-outcome (S-R-O) associations [32]. S-R-O associations are considered to be goal-directed and, despite S-R and S-R-O associations having been linked to different neuronal substrates, they share a strong relation [14] [68].

2.4.1. Instrumental learning

According to Neal *et al.* (2006), habits are "*response dispositions that are activated automatically by the context cues that co-occurred with responses during the past performance*" [69]. This statement is also in line with the Thorndike's (1911) law of exercise that states that an S-R association is strengthened every time the respective stimulus and response are paired [65].

In the past decades much has been discussed about how habits should be conceptualized and operationalized, mostly because the study of habits attracted researchers who recognize that much of our everyday life action is steered by repetition. The interest in this form of low-order behavior is supported by the key role it plays in human unconscious decision making and by its regulation through high-order processes such as cognitive control [70].

Moreover, there is common agreement that habits formation results from the acquisition of sequential and repetitive motor behaviors triggered by external or internal stimuli and consequent incremental strengthening of the S-R association [70] [71]. Consequently, the automaticity, by which actions are performed increases. This characteristic of behavior is usually followed by features such as efficiency and lack of awareness, thus allowing to allocate the attention towards something else and reduces the cognitive load of performing such action.

Goal-directed behavior, which is the other from of instrumental learning, is characterized by rapidly acquiring the encoding of the connection between action and the value of the outcome of performing that action. Therefore this form of behavior is often assumed to be the initial phase of acquisition, which is influenced by the environment. In fact the environment plays such an important role that it can elicit goal-directed behavior automatically, when the behavior associated to the activated goal became habitual. This fact supports the idea that the automaticity of behavior results of a progressive conversion of goal-directed to habitual control action [70] [72].

These two forms of behavior are not always cooperative and may actually compete for their selection and consequent evaluation of actions required to perform a choice. Nonetheless, after an extensive training, an S-R mechanism is achieved surpassing the control exerted by the participant when performing a certain action. This consequently leads to habitual actions no longer being sensitive to changes in the reward value. This reasoning is transduced in the following example: imagine you ask a participant to execute a task where the purpose is to perform a certain action in exchange to a valuable reward. If the participant continues to do it, the participant will start to perform it automatically. Then if we devaluate the reward and the participant does not present a decrease in the response to the action, we are in the presence of habit behavior [70].

Considering our study, while in the beginning the behavior might be ruled by a goal-directed approach, a considerable amount of training, through repetition of a specific action, might actually result in the automaticity of this action and in the modification of the participant's behavior towards the stimuli assigned to the action.

The process of habit development and the relationship between repetition and automaticity was, so far, addressed by Lally *et al.* (2010). In order to investigate the process of habit formation in everyday life, Lally *et al.* (2010) asked 96 participants to carry out a specific behavior always in the same context, during 12 weeks. The results showed that, for the majority of the participants, the repetition of a behavior in response to a cue was enough to develop automaticity for the behavior they performed. Moreover the authors also showed that the increase of the automaticity reached an asymptotic curve. They modeled this curve individually, since there was a considerable variation in the time each subject took to reach this asymptote [71].

Following the reasoning presented in this section, our study used Lally *et al.* (2010) study as guidance to formulate our own using the AAT, which demonstrated already findings regarding the modification of behavior through repetition [23][28][73].

2.5. The Approach-Avoidance Task

The purpose of this section is to provide more details regarding the task, since we already presented the origin and some studies explaining the importance of its use. Therefore, in this section, several factors will be highlighted.

2.5.1. Influence of the AAT on valence perception

The first finding that showed that the AAT had influence on attitudes was presented by Cacioppo *et al.* (1993). Participants were asked to rate neutrally rated Chinese ideographs while performing the AAT. The results showed that stimuli presented during arm flexion were more positively classified than the ones presented during arm extension [42].

Then, Cretenet and Dru (2004) based on the latter study and considering the aforementioned theory of Davidson, in which they assumed that the activation of the left hemispheric approach system and the right hemispheric avoidance system was, respectively, a consequence of flexion contraction of the right arm and extension contraction of the left arm.

Therefore, they tested participants on congruent conditions: extension of the left arm, simulated by positioning the respective palm on the top of one table to then exert pressure downward it, and flexion of the right arm by placing the respective palm beneath the table and then exert upward pressure. These movements were thought to activate the right hemispheric avoidance system and the left hemispheric approach system. After that the incongruent conditions were also tested (the use of the opposite palms for the specified movements). The results showed that congruent motor actions led to more positive evaluations, while the incongruent actions drove to more negative evaluations [74].

Consequently these findings supported the hypothesis that approach-avoidance reactions influence attitude formation on neutral stimuli. Although, later on, Centerbar and Clore (2006) questioned these findings because they hypothesized the pre-exposure to the stimuli could have enhanced their positivity. In fact they were right and showed that the influence of the behavior connected to the evaluation of the stimuli depended on the *a priori* valence of these stimuli. Besides that they also showed that the arm contraction did not represent a direct effect on attitude formation for neutral stimuli [75]. However, according to Ernst *et al.* (2013), this study was lacking concrete and rigorous results, which led to question the validity of what was said above [21].

Notwithstanding, Kawakami, Phills, Steele, and Dovidio (2007) found that participants who repeatedly approached black faces and avoided white ones, rated more positively the black faces than the white ones [76]. Huijding *et al.* (2009) showed that children presented higher levels of fear towards the animals they pushed away relatively to the animals they pulled [77]. In addition, more recent studies have been indicating that using the AAT for training has caused changes in the ratings of a variety of well-known stimuli such as alcoholic drinks [23], insects and spiders [73], or contamination-related objects [28]. Moreover, in the case of [26], there are actually hints that the AAT helped to reduce the participants' alcohol consumption that trained to avoid alcohol related pictures.

2.5.2. Concerns and critical issues

Currently, there are two popular versions of the AAT. The first is the joystick version, in which participants pull or push a joystick that will enhance or reduce, respectively, the picture size [78]. The other one is the manikin version where a manikin is moved on the computer screen towards the picture or away from it by pressing a button [79].

In this thesis, for practical reasons, we had to choose one of these versions. Nonetheless we knew *a priori* that this choice had some advantages as well as disadvantages and knowing these facts, here we make an allusion to the concerns and critical aspects that we thought to be worth mentioning and were already addressed in several studies [19] [78] [79].

We thought that the most important aspects to consider when comparing different versions of the same task were the sensitivity and the validity.

Regarding the first one, Krieglmeyer and Deutsch (2010) stated that the manikin version was the most sensitive of the AAT versions and gave three reasons for the differences found.

The first was: "recategorization of the required responses in other terms than the instructed response labels" [21]. For example: participants - when pushing the joystick away in the presence of a negative stimulus - might categorize the response as pushing the joystick towards the stimulus instead of away

from themselves. Nevertheless, this last hypothetical problem can be easily solved by clearly instructing participants regarding the correct arm movement they should perform. Moreover, in the feedback joystick version with its zooming effect this interpretational problem does not exist [78] [79].

The second reason concerns the distance regulation. While in the joystick version one movement simulates taking the object and the other one simulates putting it away, in the manikin version, it is the participant's position that is controlled, *i.e.*, the participant moves a stick figure towards or away from the stimulus. Thus in the first version, the hypothetical problem arises from the fact that we are moving the stimulus, while the manikin version describe natural approach or avoidance reactions of the participant by moving the stick towards or away from the desired object in order to get it or avoid it.

The last aspect that contributes to the sensitivity of the AAT is the conscious or unconscious processing of the stimuli's valence. Krieglmeyer and Deutsch (2010) stated that valent stimuli elicited approach-avoidance behaviors when subjects unintentionally evaluate stimulus valence.

Finally, to investigate the criterion-validity the authors correlated the strength of stimulus response compatibility (SRC) effects[6] for spider pictures with self-report questionnaires on fear of spiders. They concluded that the participants who presented stronger fear of spiders and rated more negatively the pictures, showed a tendentious strong behavior for avoiding these pictures. However, this only hold true for the manikin version [19].

Nevertheless, in the same year Rinck and Becker (2007), using the feedback joystick version, proved that SRC effects predicted real behavior towards spiders in a behavioral assessment test of approach and avoidance reactions [78].

For this thesis, we opted to use different variants of the feedback joystick version, in which participants were clearly instructed regarding the approach and avoidance reactions required, because it was proven that by so doing this version yielded good results. Moreover the findings reported regarding the training with the AAT, were achieved by use of the joystick version.

2.5.3. Neuronal activity during the AAT

Regarding this topic, there are only a few studies that combine neuroimaging methods and the AAT.

Actually, the majority of the studies that did it, either used functional near infrared spectroscopy (fNIRS) [31] or fMRI [80] [81] [82].

Regarding the study of Roelofs *et al.* (2009), they investigated reactions to facial expression in healthy participants and found the left lateral OFC and VLPFC to be active when performing incompatible conditions (approach angry faces and avoid happy faces) [82]. During the same conditions, Volman *et al.* (2011b) not only showed activity in bilateral VLPFC and frontal pole, but also in the fusiform gyrus, left supramarginal and inferior parietal gyrus [81].

Moreover, they also found the expected pattern of faster RTs for compatible reactions, when the subjects evaluated the pictures according to the valence of the stimulus.

Regarding the study of Lena *et al.* (2013), the authors assessed prefrontal activity in healthy subjects using fNIRS during the performance of the joystick version of the AAT and found that incompatible

[6]The increase of RTs in incompatible compared to compatible conditions.

reactions (approach negative and avoid positive) triggered higher activation of the DLPFC compared to compatible ones. Moreover, in a second study, now using alcohol and non-alcohol related stimuli, they found higher activation of the left anterior lateral orbitofrontal cortex when approaching alcohol related stimuli compared to avoidance reactions of the same stimuli [31].

Finally, in [80] the authors examined 37 subjects of which 20 were alcohol dependent and 17 were healthy subjects using fMRI and the joystick version of the AAT. In this study, the subjects pushed and pulled pictorial cues of alcohol and soft-drink beverages, according to a content irrelevant feature of the cue (landscape/portrait).

The results showed that the alcohol dependent group presented stronger behavioral approach tendencies towards alcohol cues compared to the soft-drink pictures. Furthermore the fMRI results showed larger activation of the nucleus accumbens and medial prefrontal cortex, regions involved in reward and motivational processing, in the case of the critical fMRI contrast defined for alcohol-approach: (approach alcohol > avoid alcohol) > (approach soft drink > avoid soft drink). It was also found a positive correlation between alcohol craving scores and activity in the amygdala for the approach alcohol contrast, in alcohol dependent patients [80].

To sum up, it is important to emphasize that all these findings are complementary and give first insights regarding the neuronal substrates that are possibly active when performing the routines we present in the following chapter.

3

Methods

Contents

In this chapter, an approach which connects computational models with behavioral findings is described.

Firstly, we describe the different versions of the AAT, from which the behavioral data was acquired, including details on stimuli used and a brief description of the protocol used in this study. Then a brief description of how the data was acquired and pre-processed is given.

Secondly we introduce the mixed-effects models and how we employed them. After that a detailed description of our computational model is provided along with remarks to take into consideration, and a set of *a priori* hypotheses is defined.

Then we introduce how the parameters of our model were estimated and refer the processes and solutions which we found to the problems that we faced.

Finally, we elucidate how we compared the different tested models and how the best models were selected.

3.1. Implemented versions of the AAT

In a general routine of the AAT, participants have either to approach or to avoid stimuli presented on a computer screen, according to a specific instruction. In the version that was implemented in the current study, the joystick version with feedback, participants had to pull or push a joystick, which increased or reduced the picture size, respectively. This effect, which was called by Rinck and Becker (2007) as zooming effect, allowed participants to better perceive the idea that by pulling the joystick they were approaching the stimulus and by pushing the joystick they were avoiding it [78].

In order to assess different behavioral aspects when performing the AAT, three different versions were created through modification of the original version of the joystick version with feedback (cf. sub-sections 3.1.1; 3.1.2 and 3.1.3).

To perform each of the different versions, participants were seated in a chair with a pillow in a viewing distance of approximately 40 cm from the computer screen of an ASUS K450J Notebook and reacted via a joystick Logitech, Extreme 3D Pro with their dominant hand. Moreover, they were told to place the joystick in a way they felt comfortable with. The reason for seating participants on a pillow during the whole week is explained in sub-section 3.1.7.

Participants also went through a small training in order to perceive the joystick sensibility and the instructions that they would receive in the following routines. This was performed because there were several participants that said that they never had worked with a joystick. After this, before they started any routine, they were told “to be as fast and accurate as possible” and to place their non-dominant hand over the base of the joystick to better perform the routines.

3.1.1. Training version

The training version consisted of a routine where participants of different groups trained different conditions. Participants of the negative group were trained to approach negative pictures and to avoid neutral ones, while participants of the positive group were trained to approach neutral pictures and to avoid positive ones. Each condition was trained for 60 trials (30 trials per image in each condition, since participants trained two images per condition), yielding a total of 120 trials.

Therefore there were two types of trials: the ones where participants had to approach the stimuli and the ones where they had to avoid the stimuli. The general structure of both trials is depicted in figure 3.1.

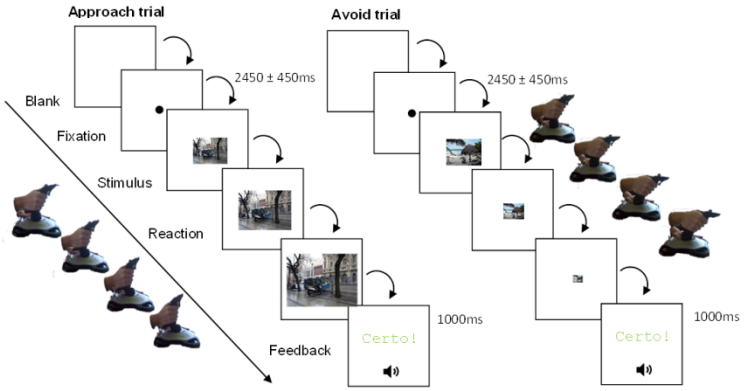


Figure 3.1: Schematics of a typical trial in the task composed of 5 events: fixation, stimulus, reaction, feedback and blank. Temporal differences between each event are also depicted for both types of trials. The sonorous icon corresponds to the sound that accompanied the visual feedback.

At the beginning of each trial a fixation point was presented on the center of the screen for 1000 milliseconds (ms) (fixation event). This was followed by a randomly jittered interval between 1000 and 1900 ms (waiting event), during which the fixation point was maintained. During these two events the participant was notified if the joystick was or not in the correct position. If not, the following message appeared on the screen "Endireita o joystick" (Put the joystick into the initial position). Next, the stimulus appeared on the screen for an unlimited time (stimulus event), during which the participant had either to pull (*i.e.*, approach) or to push (*i.e.*, avoid) the joystick and consequently the stimulus.

For stimulus presentation, a picture of medium size was displayed in the center of the screen (resolution 400 x 300 pixels, size 8 x 10.5 cm). The zooming effect was generated by drawing different picture sizes in relation to the position of the joystick. In fact, for each direction, 500 pictures were drawn. The movement of the joystick was constrained between angles that correspond to a joystick position of -1 (joystick totally pulled) and +1 (joystick totally pushed) and we created a vector of 500 points that determined the size of the picture displayed at each moment (according to the current position of the joystick). This was accomplished by summing or subtracting to the original dimensions a weighted quantity based on the distance of the joystick to the original position (called rest position). Irrespective of whether the joystick was moved in the correct or wrong direction, the picture disappeared as soon as the position of the joystick reached either 1 or -1. Motions to the left and right side had no effects on the picture being shown.

After stimulus presentation, feedback regarding the participant's action was shown. This remained on the screen for 1000 ms, and it was either a green "Certo!" ("Right!"), accompanied by a metallic and pleasant sound after correct responses, or a red "Errado!" ("Wrong!"), accompanied by an unpleasant buzzer sound.

After completing the block of 120 trials, we asked participants to rate the pictures they were presented with, in terms of pleasantness. In order to execute this, participants used an external mouse to select a square in a bar with 201 squares (that coded from -100, for "the most unpleasant thing I have ever imagined", to 100, for "the most pleasant thing I have ever imagined", and had 0 as "neutral"). For this,

the following question was phrased: “How unpleasant or pleasant is this picture to you?”. The mouse cursor was represented by a single dot, which was always presented at the center of the screen and consequently on the center of the picture (figure 3.2).

Figure 3.2: General scheme structure used for participants to rate the pictures.

Besides the provided verbal instructions (cf. section 3.1), participants were instructed according to figure 3.3.

Figure 3.3: General structure of the instructions provided to participants. The instructions depicted were specific for the positive group’s participants. The first sentence says: “When you see these pictures, approach them by pulling the joystick”; the second says: “When you see these pictures, avoid them by pushing the joystick”; the last sentence says: “To begin the training, press A on the keyboard”.

With this routine we are training people to react to positive and negative stimuli in an opposite way of the automatic tendency they have towards these stimuli. Moreover we also had a neutral condition for each group, because we found literature that used the neutral condition as a control condition [28].

This routine was based on the one that Eberl *et al.* (2013) had used to train alcohol-dependent subjects to avoid alcohol related stimuli and which had resulted in a tendency for reduced relapse probability after one year [26].

3.1.2. Assessment version

The assessment version was created in order to capture the effects of the unintentional valence processing of participants. In this version, which consisted in a routine of 192 trials in total, we presented 16 pictures to participants that were grouped in 3 different categories (cf. sub-section 3.1.5). Each picture was shown 12 times (6 for one direction and 6 for the other). This routine was performed on the first day before the training, and on the last day, after the last training session.

The trials' structure of this routine was similar to the one presented in sub-section 3.1.1 (cf. figure 3.1). Nonetheless, some modifications were performed. Firstly, the stimuli presentation was done along with the instruction of which action the participant should perform (figure 3.4a and figure 3.4b).

Therefore, participants were instructed by the arrows as depicted below: They had been designed in order to provide a clear instruction and to avoid any interference with the content of the picture presented.

Figure 3.4: General instructions and stimuli provided to participants in the assessment version. (a) Indicates that the participant should avoid the picture by pushing the joystick and (b) indicates that the participant should approach the picture by pulling the joystick.

Secondly, even though participants only received feedback when they wrongly performed one trial, the time intervals between each event depicted in figure 3.1 were kept.

Thirdly, before they started the routine a written instruction appeared on the screen phrasing: "Pull or push the joystick according to the arrow's direction".

Fourthly, on the first day, the rating of the presented pictures was performed before they started this routine, while in the last day the rating of the pictures was performed after completing the task.

The ratings were executed as above mentioned because we wanted to verify if there were any differences in how participants perceived the pictures before and after the training. Besides using this routine to investigate the effects on the trained pictures, we also aimed to use it to inspect if the effects of the training were generalized for pictures similar to the ones participants had trained. We analyzed this because we expected participants who trained a specific condition (*e.g. approach negative*), to be faster when performing the trial where an arrow indicated to approach similar negative picture, than when performing a trial where the arrow indicated otherwise.

The arrows were a novelty and this version was based on a similar routine where participants were instructed by the shape of the picture's frame presented [83]. We decided to use the arrows because we expected them to reduce the WM load, so that participants really associated the content of the pictures with the trained reactions.

3.1.3. Arrow version

The arrow version consisted of a total of 20 trials (10 for each direction) and was performed in the first day, before the assessment version.

Regarding the trial's structure of this version, the difference to the trial's structure presented in sub-section 3.1.1 (figure 3.1) was the stimuli presented to participants. Figure 3.5 depicts how participants were instructed in this version of the task and the presented pictures.

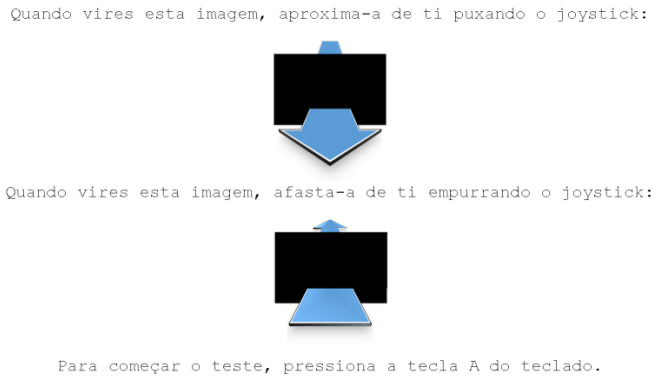


Figure 3.5: General structure of the instruction provided to participants, in the arrow version. The first sentence says: "When you see this picture, approach it by pulling the joystick"; the second says: "When you see this picture, avoid it by pushing the joystick"; the last sentence says: "To begin the training, press A on the keyboard".

This version was designed in order to assess participant's motor biases when using the joystick. Therefore we used a black rectangle, with the size of the pictures used in the other versions, to not interfere with the measured RTs. The data acquired from this part of the experiment could then be used to estimate a ratio between avoidance and approach reactions and identify which participants had a bias for one specific movement. These would have been of importance when, at group level, these biases would not have canceled out (cf. sub-section 4.2.1).

3.1.4. Computational implementation

The task versions described above were developed in MATLAB version R2012b, using Psychophysics Toolbox Version 3 (PTB-3). This toolbox consists of a free set of MATLAB functions that allow us to fully control the hardware for precise stimulus display.

The implemented algorithm was very flexible, which allowed us to design the different versions of the AAT. The main part of the algorithm is based on a joystick version of the AAT with feedback.

Details on this programming are given in appendix A1.

3.1.5. Stimuli

For this study we decided to use 16 pictures from three different categories: positive, negative and neutral. Four pictures were assigned to each of the first two categories, while the latter contained 8, as the neutral pictures were split into two: the ones to use along with the positive pictures and the ones to use along with the negative pictures. This division was done in order to have two groups of participants: the positive group, that trained positive and neutral pictures, and the negative group, that trained negative and neutral pictures. More specifically, we were interested in training participants on two conditions: approaching the negative pictures and avoiding the positive ones. To compare these incongruent conditions with a control condition, a neutral condition with the opposite movement was added for each group.

Considering the fact that these tasks will be applied in OCD patients in the future, the choice of the pictures was based on pictures that could elicit strong automatic avoidance reactions (similar to the pathologically enhanced avoidance tendencies which can be found in OCD patients for specific stimuli). Thereby, the pictures had to be salient for the tested healthy participants. Therefore, the chosen negative pictures depicted very dirty toilets. Regarding the neutral pictures for the negative group, we

decided to use neutral kitchens because both pictures types depicted scenes within a building, within a living area that can be found in every household.

Regarding the positive pictures, we chose pictures related to vacation scenes (similar to the one depicted in figure 3.2), because we considered this stimuli to be doubtlessly positive. Consequently, they should elicit quite strong approach reactions. Relatively to the neutral pictures for this group, we chose pictures related to usual outside scenarios in a city.

The pictures used were downloaded from a very large database generated by Microsoft [84]. This database was composed by 44 504 pictures that were analyzed to select candidate pictures to be used (negative: 154718, 264964; neutral: 641, 10114, 10844, 18366, 20979, 26204, 27285, 28682; positive: 348896, 490337, 556420). We also would like to refer that several links were analyzed with pictures that were tested for valence properties. However, excluding pictures from the International Affective Picture System (IAPS), none of the analyzed databases had pictures of our interest. Even the database of the IAPS [85] only had some useful pictures (negative: 9300, 9320). Nonetheless, since we needed many similar pictures to check for generalization effects, we had to resort to some pictures available on the internet. Thereby, the pictures we downloaded were carefully chosen in order to avoid copyright issues.

Finally, regarding the characteristics of the pictures, after some treatment using the Paint.NET v4.0.4 software, all of them end up having the same dimensions (400x300 pixels).

3.1.6. Protocol

In order to guarantee that all participants were under similar circumstances, we assured that they executed all the procedures by the same order and in a place without any external distractions that could influence or compete with the attention we required participants to pay when performing the routines explained in the sub-sections 3.1.1, 3.1.2 and 3.1.3.

Therefore, we created an initial protocol, in which we registered and oriented all the procedures each participant was subjected to. This initial protocol was used on a first sample of participants that we acquired to verify, if all the implemented routines provided the expected outcomes and, if, on the other hand, if there was to change any detail of them.

A short version of the final protocol and duration of each activity performed is presented in the table 3.1.

<table>
<tr><th>1st Day</th><th>2nd Day</th><th>3rd Day</th><th>4th Day</th><th>5th Day</th></tr>
<tr><td>Questionnaires (20 minutes)</td><td rowspan="6">Training (10 to 15 minutes)</td><td rowspan="6">Training (10 to 15 minutes)</td><td rowspan="6">Training (10 to 15 minutes)</td><td rowspan="2">Training (10 to 15 minutes)</td></tr>
<tr><td>Pre-train (3 to 5 minutes)</td></tr>
<tr><td>Arrow (3 to 5 minutes)</td><td rowspan="2">Assessment (15 to 20 minutes)</td></tr>
<tr><td>Assessment (15 to 20 minutes)</td></tr>
<tr><td>Training (10 to 15 minutes)</td><td>Questionnaires (5 minutes)</td></tr>
<tr><td>Questionnaires (10 minutes)</td><td>Practical Test (3 to 5 minutes)</td></tr>
</table>

Table 3.1: Summary of the activities participants were asked to perform and respective durations.

From table 3.1, it can be inferred that participants were subjected to 5 consecutive days of training. The first training session was preceded by one pre-assessment with arrow version and one test with assessment version. Then, after the last day of training, participants were subjected again to the assessment version of the task and to a practical test.

As last remark we would like to refer that the data of a specific participant was collected at the same period of the day. The questionnaires that participants had to fill in were not analyzed for this thesis, but will be part of further analyses. These questionnaires were used to assess the mood and general personality traits that might influence the reactions in the AAT.

3.1.7. Practical test

In order to have another measure of the effects of the training, a practical test was designed. This test consisted on measuring the time participants took to sit on a pillow with a modified cover. This modified cover had two different pictures printed on it: one related to positive stimuli (a paradisiac beach, figure 3.6a) and the other related to negative stimuli (a very dirt toilet, figure 3.6b).

Figure 3.6: Pictures depicted on the cover of the pillow where participants sit when performing the practical test: (a) shows the positive picture, (b) the negative picture.

In order to avoid confounds of seeing the pillow for the first time, we asked participants to sit on a normal pillow during the whole training week. Before the practical test, participants were instructed to fill in one questionnaire. Then, participants were asked to sit on a chair that was under the table, as usually when they arrived. The experimenter pressed one key (the space bar) to count the time from the moment participants started to pull the chair away from the table until they sat on it.

Meanwhile, on the computer screen, an instruction appeared saying: “Depois de te sentares pressiona a barra de espaço” (“After sitting on the chair please press the space bar”). At this moment, the time was stopped and the participants were instructed to rate pictures from three different categories: positive, negative and neutral, using the method explained in sub-section 3.1.1 (cf. figure 3.2).

These pictures were obtained from the database generated by Microsoft [84] and from IAPS (positive pictures: 4689, 8501; negative pictures: 1280, 1525, 6250; neutral pictures: 7010, 7175, 7090) [85].

The time was acquired using a simple, but very accurate MATLAB routine.

3.2. Behavioral datasets

As stated in section 3.1, two samples of data were acquired.

3.2.1. First sample

This dataset was constituted by 10 healthy adult participants (5 women and 5 men) with ages between 22 and 25 years, mean = 23.3 and standard deviation (SD) = 0.949 years. The participants were pseudo-randomly distributed to the two groups and the negative group was constituted by 3 women and 2 men whereas the positive one contained 2 women and 3 men. Within each group, participants trained the same conditions and the same images.

This sample did not execute the neither arrows version nor the practical test, but performed a short version of the training and the assessment version, which had in total 60 and 128 trials, respectively.

Nonetheless all subjects performed structurally the same tasks, in which the sequence of pictures presentation and waiting times was randomized across subjects from each group in each session.

All the participants were friends of the experimenter and performed the tasks voluntarily, after providing a written informed consent, regarding the use of their data. This consent was approved by the local Ethics Committee for the Health Care of the University of Lisbon.

The participants reported their dominant hand as being the right one and even though some participants never had had any kind of experience working with joysticks, they also reported there was not any problem with it. Some participants also reported that moving the joystick in one direction was easier than moving into the opposite one. However these effects were dispersed, since nearly half of the sample indicated one direction and the other the opposite. Nevertheless, for the second sample, the arrows version was created in order to then analyze those effects.

We report the results of this first sample in the appendices (cf. appendix A2) due to the fact that they were pilot data, but not part of the rigorous statistical main analysis. Still, these results were important, since they significantly contributed to the development of the final task versions. More specifically, after this first experiment, the number of trials was adjusted for the second, main experiment, because these results suggested that the learning curves could have decreased more if the training period was extended.

3.2.2. Second sample

This dataset was composed by 36 adult participants (18 women and 18 men) whose ages varied from 19 to 29 years, with mean = 23.056 and SD = 1.754 years. Due to hardware limitations (just one joystick and one computer) the acquisition of data was done during three weeks. Once again the participants were pseudo-randomly distributed to the negative and positive group, to assure that both groups were balanced in gender, *i.e.*, each group was composed by 9 women and 9 men.

In this second sample, all 16 pictures were used in the training version of the AAT. Nonetheless, considering that each participant trained 4 pictures (two of each condition) and there were 4 pictures per condition, we had 6 possible combinations of pictures per condition. Therefore, we randomly

distributed these 6 possible combinations across subjects within each group, ensuring that each combination of pictures was trained by the same number of participants.

Every participant performed the same tasks, following the protocol that was described in sub-section 3.1.6, in which the sequence of pictures presentation and waiting times was randomized for each subject from each group in each session.

All the participants were acquaintances of the experimenter and performed the tasks voluntarily. Each of them provided a written informed consent, regarding the use of their data, which was approved by the local Ethics Committee for the Health Care of the University of Lisbon.

In this dataset, two participants reported to be left-handed and, therefore, performed the tasks with their left hand. They, as all the other participants, were inquired regarding the comfortability of the joystick to execute the several routines. None reported to be uncomfortable when handling it.

Regarding the analysis of this sample, it is important to refer that the percentage of eliminated observations due to errors and outliers in each subject was not high (the maximum percentage of excluded trials (cf. section 3.3) was 7.8% and the mean percentage was 3.28%). For that reason, all participants were included in the analysis.

3.3. Data pre-processing and outliers' exclusion criteria

The first step was to analyze if participants performed the task as they were asked to do. According to prior analyses of reaction times acquired during the solving of behavioral tasks [23] [26] [27] [37], the wrongly performed trials were not considered for the analysis. Besides that, we also did not consider the trials where participants' initial movement was the opposite of what was instructed.

After this, we also analyzed the data for additional outliers. In fact, in the analysis of RTs, this is a very difficult problem to deal with. Outlier RTs might be generated by processes that do not correspond to those we want to study and might have several underlying processes such as fast guesses, subject's estimate of the usual time to respond, subject's inattention or fatigue [86] [87] [88].

According to Ratcliff (1993) there are two main types of outliers when dealing with RTs: the short and the long RTs. In our case the first ones were less frequent, since there is a natural limit in speed under which the joystick movements cannot be performed, while the second ones posed a real problem.

When analyzing the literature related to the analysis of RTs, several criteria for pre-processing of correct RTs were found [86] [87] [88]. Although, there are some statistical tests to spot outliers, these methods are usually not appropriate for RTs distributions [86]. Therefore, we decided to use a mixture of the criteria found. First we used a specific cut-off at 200 ms, because we did not expect meaningful RTs to be faster than this for our task. Additionally, genuine RTs cannot go below 100 ms, which is the minimum time needed for physiological processes underlying stimulus perception to very simple motor responses, such as a button press [88].Therefore, 200 ms seemed a reasonable cut-off since both these processes are needed for the AAT execution with the more complex joystick movement. The other criterion used was the cut-off at 3 times the interquartile range above the third quartile. This prevented to eliminate meaningful information as might happen with the previously used the conservative or the lenient criteria, which are cut-offs at one and two SD above the mean, respectively [87]. Also, the use of a cut-off at some number of standard deviations above the mean is not recommended as both the

average and the SD might be inflated by the abnormal measures. Moreover, according to Miller (1991), the use of means to eliminate outliers in studies that acquire RTs is not a good practice, since in most of the cases the RTs do not follow a normal, nor even symmetric, distribution. Therefore, it would be unlikely that such a method would be suitable to exclude observations from the upper tail [87] [88], as we wished to do.

After having decided which criteria to use, the question was how to apply them. Since there was a clear inter-subject variability (cf. appendix A4) we could not perform this analysis at group level. Therefore, we decided to apply the chosen criteria at an individual level.

Nonetheless, to support the decision regarding the used outlier criteria, an exploratory data analysis regarding the RTs was done. This was performed through estimation of the density function that best characterized the RTs for each subject. To do so, we resorted to *histDist* function in R, which allowed us to have visual overview while providing an output to which we could apply distribution test functions such as the *Kolmogorov-Smirnov test*[7] and then infer about the results. In fact, the *histDist* function provides a histogram estimate of the density function, which is one of the most commonly used estimates in RTs studies [86] [89].

It is important to notice that this analysis had also provided a hint to identify which method we should use in order to best estimate the parameters of our computational models (cf. section 4.1).

3.4. Mixed-Effects models

A mixed model is similar in many ways to a linear model, because it describes the effects of at least one predictor on the variable of interest [90].

However, a mixed-effects model (or just mixed model) arises from the incorporation of both fixed effects, that are parameters which tell how population means differ between any set of treatments, and random effects which in turn are parameters representing the general variability among subjects [91].

Since our data was correlated due to grouping of subjects, we decided to use this method as it is a very powerful and flexible tool for the analysis of such situations, by allowing to explicitly model a large variety of correlation patterns.

Another interesting feature of these models is that the simultaneous use of fixed and random effects can be thought as hierarchical modeling, since we have one level for subject's effects (random effects) and another one for measurements between subjects (fixed effects).

This family of models is usually represented in terms of three random variables: a *q*-dimensional vector of fixed effects ($\boldsymbol{\beta}$), a *q*-dimensional vector of random effects ($\boldsymbol{b}$) and an *n*-dimensional response vector ($\boldsymbol{y}$). The latter has the values $\boldsymbol{y}$ which we observe, while the values $\boldsymbol{b}$ and $\boldsymbol{\beta}$ are the ones we want to estimate. To do that and make inferences about them we use predictors.

A linear mixed model generally follows equation 3.1.

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{b} + \boldsymbol{\varepsilon} \tag{3.1}$$

[7] To use this test we assumed that the parameters estimated from the raw data were the real parameters of the data.

Where $\boldsymbol{\varepsilon}$ is an unknown vector of random errors, and X and Z are design matrices that relate the unknown vectors $\boldsymbol{\beta}$ and $\boldsymbol{b}$ to the vector of observations $\boldsymbol{y}$.

These models were used to fit the behavioral data acquired from the training, and random effects were considered for both the intercept and the slope. The random intercept was estimated as the variance of the individual intercepts around the common intercept of each group. The random slope was done by additionally estimating the variance of individual slopes around the common slope for the respective group of subjects.

This method was also used to analyze the behavioral data acquired from the other AAT routines. There would have been other methods derived from classical statistics such as analysis of variances (ANOVAS) (*e.g.* Repeated Measures ANOVAS) to perform those analyses, but since mixed models were shown to be more sensitive, we applied them [92] [93]. These studies showed enhanced sensitivity of mixed models due to their ability to model nonlinear, individual characteristics.

During the last decade, this approach has become more and more popular [94] since it allows analyzing dynamic phenomena in a more flexible way. Compared to the classical ANOVAS, this approach allows to characterize group and individual behavior patterns in a formal way, providing both group and individual differences and incorporating additional covariates [92] [93].

In order to build the different models (cf. table 4.2) which we wanted to test, we used the software for statistical computing R (R version 3.1.1) and more specifically the functions available in the lme4 package.

3.5. Novel computational models

As we mentioned in sub-section 1.4.2 nothing had been previously done regarding the computational modeling of subject's behavior, when performing the training version of the AAT. Therefore we wanted to model subjects' behavior, *i.e.*, to fit the collected behavioral data, through novel computational models that captured the influences of some psychological and cognitive processes which we thought to be involved when performing the AAT. These will be described next.

To fit individual subject's behavior, the developed models use several information such as the actions which the participant performed, the instructions that were given, the type of stimuli, the participant's rating of the stimuli and the RTs associated to each trial.

So, to model subjects' behavior when performing the training version of the AAT, we aimed to model the preference of the subject $w_t(s_i, a_j)$ for executing a certain action, a_j, when presented with one specific stimulus, s_i at trial t. The preferences were modelled according to three specific influences: habit learning (h), pavlovian biases (p), and cognitive control (c), in agreement with the concepts explained in chapter 2 and according to equation 3.2.

$$w_t(s_i, a_j) = h_t(s_i, a_j) + p(s_i, a_j) + c(s_i, a_j) \tag{3.2}$$

Regarding the habit learning component, we assumed that the training of specific conditions, *i.e.*, repetitions would lead to the learning of new S-R associations between the stimulus s_i and the respective instructed action a_j. This was done in agreement with Thorndike's law of exercise (enunciated and explained in chapter 2) and Neal *et al.* (2006), which brought to light evidence concerning the role of

repetition in habit learning [95]. Moreover, several studies consistently reported that training approach and avoidance tendencies is an effective way to regulate impulsive and emotional behavior. For instance, Eder *et al.* (2011) showed that the incongruent movements become linearly faster with practice [96]; Wiers *et al.* (2013) showed that training drug users to avoid drug cues, in a paradigm very similar to ours, resulted in learning to avoid this condition [97].

Therefore this component was modelled by equation 3.3.

$$h_{t+1}(s_i, a_j) = (1 + \alpha) \times h_t(s_i, a_j) + \beta \times \frac{0.5}{1+T(s_i)},\ \alpha \geq 0,\ \beta \geq 0 \tag{3.3}$$

Where α and β are the multiplicative and additive learning rates, respectively.

The term that is multiplied by β was implemented to simulate a decay on this parameter that results from the experience of repeatedly observing the stimulus s_i [98]. Thus, $T(s_i)$ is the number of past observations of stimulus s_i.

The multiplicative learning rate, on the other hand, tried to capture the occurrence of Hebbian learning, and was based on the three factor Hebbian rule, which states that learning of a *stimulus-response* association is dependent on presynaptic activation, on dopamine, and on the post-synaptic activation in the striatum [99]. Since no prediction errors are computed in our task, we used a slightly different implementation of this rule, in order to model the dopamine-mediated strengthening of synaptic associations caused by the cumulative experience of the *stimulus-response* pairing.

Relatively to their influence on the preference, the higher the values of the learning rates were the faster the participant would learn the S-R association.

Concerning the Pavlovian component we assumed it to be responsible for the subjects' innate bias towards congruent or incongruent reactions, for a specific stimulus s_i. Thus, this component was modeled by the product of two parcels: the valence of each stimulus s_i ($v(s_i)$) and a subject-specific Pavlovian parameter (π).

This component is governed by appetitive or aversive valence associated responses, and contributes to impulsive behavior [100]. Its modeling was also supported by several studies [29] [31] [43] [56] which suggested that positively valenced stimuli facilitate actions usually linked to approach reactions, while negatively valenced stimuli facilitate actions that lead to avoidance reactions [101].

Thereby equation 3.4 which ruled this component is given by:

$$p(s_i, a_j) = \begin{cases} -\pi \cdot v(s_i),\ if\ a_j = avoid \\ \pi \cdot v(s_i),\ if\ a_j = approach. \end{cases} \tag{3.4}$$

Since we also wanted to take into account the sensitivity of the subject towards the stimuli, we allowed the strength of this effect to vary according to subject's evaluation. Thus, $v(s_i)$ is driven by the rating the subject attributed to that stimulus before initiating the training period. Therefore, the influence of this component on the preference $w_t(s_i, a_j)$, will be big as $|v(s_i)|$.

Therefore, it is clear that congruent reactions, meaning approaching positive or avoiding negative stimuli, are facilitated, while the incongruent ones are hindered by positive values of the Pavlovian parameter.

The cognitive control component tried to model the cognitive effort people had to engage when they were faced with a situation where they were asked to act incongruently. Thus, this component refers to the ability of flexibly allocating mental resources; in this specific case, to guide actions in the presence of competing automatic reaction tendencies. Besides this, under such demand, cognitive control serves to direct the processing of goal-relevant information and to schedule actions in order to minimize conflicts between potential responses. Moreover, this component has been associated with a neuronal network composed by brain regions such as the PFC, the DLPFC, the ACC and others [31] [61] (for further details on this see chapter 2).

Notwithstanding, the fact that the system DLPFC-ACC had been implicated in tasks demanding the overcome of a prepotent response tendency [61], in this initial phase, we were interested more specifically to model the degree of cognitive control engaged by subjects. Therefore, this component was described by equation 3.5.

$$c(s_i, a_j) = \begin{cases} C, \ if \ [(v(s_i) < 0) \wedge (a_j = approach) \wedge (instruction = approach)] \vee \\ \qquad [(v(s_i) > 0) \wedge (a_j = avoid) \wedge (instruction = avoid)] \\ 0, \ otherwise \end{cases} \tag{3.5}$$

Where C represents the degree to which a participant engages cognitive control ($C \geq 0$).

Having explained the components which we thought to influence the preference $w_t(s_i, a_j)$ we had then to transduce these psychological processes into behavioral measures. To perform that we resorted to Piéron's Law.

This is a psychophysical regularity in signal detection tasks, which states that RTs decrease according to a power law with the increase of stimulus intensity (equation 3.6).

$$RT = \alpha I^{-\beta} + \gamma \tag{3.6}$$

Where α and β are free parameters that determine the slope and amplitude of the function and γ is an intercept [102].

Regarding equation 3.6, it is important to notice that RT is a sum of two factors to which can be assigned different interpretations. According to van Maanen *et al.* (2012) while γ accounts for the non-detection related processes of the response time (then named non-decision time parameter), the $\alpha I^{-\beta}$ factor is associated to the detection time (then called decision time) which is related to the rate of accrual of a given stimulus intensity [103].

More interestingly, although this law was formulated to describe an effect of stimulus intensity, for the past decades it was proven that this relationship also holds true for two-alternative choice tasks [102] [103]. One example of that is the study of Stafford *et al*. (2011) that showed that RTs decrease as a power law in the Stroop color naming task with the luminance of the color dimension [104]. This finding led van Maanen *et al.* (2012) to hypothesize that Piéron's Law could be related to a more general notion of discriminability in decision-making, *i.e.*, the stimulus intensity could be seen as a interfering factor on the decision, since a low intensity stimulus discriminates poorly between two alternative responses. In fact, van Maanen *et al.* (2012) did not only extend this relationship to perceptual two-choice decision making tasks, but also proved that Piéron's Law was still applicable when the discriminability of two competing choices was manipulated (for further details see [102] [103]).

Considering these findings, we transformed the preferences into predicted RTs (r_t) through the equation 3.7.

$$r_t = [w_t(s_i, a_{instructed}) - w_t(s_i, a_{non-instructed}) + k]^{-D} + E \quad (3.7)$$

Where E represents the non-decision time, D controls the decrease in RTs as the relative preference for the instructed response increases and k is a constant which was added due to numerical stability concerns of the model. The value computed for k was performed considering the worst-case scenario, that is, the case where the relative preference was zero and the parameter D had a value equal to its upper bound.

Thus, the higher the relative preference, the lower would be the predicted RTs, according to equation 3.7. Moreover, as desired, following this set of equations, until the habit learning component became considerable, the preference, and consequently the predicted RTs, would be mainly ruled by the interaction between the cognitive control and the Pavlovian components (cf. equation 3.2).

3.5.1. Remarks

An important note is that we had 6 parameters to estimate, therefore we found it useful to increase the number of trials participants had to perform, in order to have sufficient data points for the routine of optimization to accomplish a good fitting without overfitting[8], which would results in misleading conclusions. Besides this, we also found some multicollinearity problems, since the free parameter D influences the four parameters used to compute the preferences $w_t(s_i, a_j)$ (cf. equation 3.7).

Before the fitting process starts, the model defines initial conditions for the parameters, which, given the model design, have to be constrained. Otherwise the optimization routine would run into numerical problems all the time and to inconclusive results.

According to the previous reasoning, the main constraint which was imposed was that the cognitive control component should be higher than the Pavlovian component (equation 3.8).

$$C > |\pi \cdot v(s_i)| \quad (3.8)$$

With this we guaranteed that there were not any negative preferences (since both learning rates were constrained to be non-negative), which, if allowed to exist, would lead to the prediction of imaginary RTs. Moreover both the Pavlovian and cognitive components were modeled as constant variables for simplification that actually was proven to be reasonable through Bayesian comparison approaches (cf. section 4.3).

Additionally, we though that fitting models to data points obtained by performing a moving average on the subjects' RTs would provide complementary information of great interest. This because the undesired processes that we could capture by fitting spurious RTs (cf. section 3.3), could subsist even after the execution of our data pre-processing routine.

A moving average is mathematically transduced by equation 3.9.

[8] Statistical problem that occurs when a model is excessively complex, such as having too many parameters. Leads to poor predictive performance.

$$ma_i = \frac{1}{n}\sum_{j=i-\frac{n-1}{2}}^{i+\frac{n-1}{2}} x_j \quad (3.9)$$

Where x is the vector of the original RTs and n represents an odd window size, *i.e.*, the number of points we considered for the computation of the moving average.

Although the moving average method might not be completely unproblematic when dealing with frequencies (because of its poor ability to separate one band frequency from another), in our specific case that was not a problem.

As a matter of fact the real problem faced when applying this method was to take into consideration that we could not blindly apply it to one subject's data, since the RTs were assigned to specific pictures of a certain condition. Besides this, using this method could also mean loss of information, especially because the initial RTs would be averaged out with the following ones that normally would already be lower, interfering with the original shape of the raw data.

Despite the problems pointed out above, we still found it useful to apply this method. Therefore, the main problem of loss of information concerned the first data points since they did not have any previous information to be averaged with. One possible solution was to use the points themselves until we reached a point that allowed the use of this standard filter in terms of the defined window size. However, this was not considered since it could increase even more the discrepancy between the first points and the following ones. Another possible solution was to use half of the window points to apply this method, for example, if we considered a 5 point window, the first point would be averaged by the next two points and itself. Considering the circumstances, this appeared as a good solution since it would not interfere too much with the original shape and it would not create a gap as big such as the other hypothesis would.

3.5.2. Implementation

The framework from which the computational models were created was developed using MATLAB version R2012b, through the creation of a function that was responsible for three main tasks:

- Reading and processing of the subject's output data.
- Estimating the best set of parameters, and consequently the best set of predicted RTs, to fit the behavioral data (*i.e.*, the RTs) acquired from each subject.
- Storing the predicted data concerning each subject, including the observed and predicted RTs and the estimates of the variance of the distribution which the RTs were assumed to follow during the fitting routine (cf. sub-section 3.6.2).

For the second task we developed two functions that were very similar. While the first one was responsible for the process of parameter estimation and was used by the optimization routine, the second was used not only to compute the predicted RTs with the best set of parameters found by the previous function, but also to punctually estimate the sub-optimal variance of the predicted behavioral data (as described in sub-section 3.6.1).

These two functions were responsible for the computation of the preference at each trial for a specific action ($w_t(s_I, a_j)$) and contained another three auxiliary functions that helped to calculate the three components that influenced $w_t(s_I, a_j)$.

Additionally, other functions were implemented in order to pre-process the behavioral data, to compute the initial conditions (*i.e.*, to compute a set of initial values of the parameters used for the optimization routine to start with at each iteration), and to provide graphical representations of the output from the model.

Moreover, it was also implemented a specific set of functions that formatted the output from the data fitting function, allowing its subsequent use in the model comparison toolbox (cf. sub-section 3.7.2).

Besides this, in a final part of this thesis, due to the fact that this routine was taking too much time to be completed, we decided to implement an analogous model in STAN.

Shortly, STAN is a probabilistic programming language which compiles the model we implemented to C++ code, allowing to use optimization routines to maximize objective functions. Besides, STAN can also be used for Bayesian modeling and inference using sampling algorithms to obtain posterior simulation of a specific model and data [105].

Therefore some auxiliary functions were implemented, in the software for statistical computing R (R version 3.1.1), to adapt and process the behavioral data to then be fitted, using the function optimizing (that is described in the section below) of the *rstan* package available for R. Although this process usually takes a while, the optimization process was more efficient and less time consuming.

3.5.3. *A priori* hypotheses

Considering the study and the aims of this thesis, we first predicted that the RTs would not be normally distributed.

Secondly, we expected this model to partially capture some of the psychological and cognitive processes of interest, meaning that it would better describe the behavioral data than a chance model, even though this was just a preliminary experimental model.

Thirdly, given the fact that there were pictures which elicited stronger reactions, it could be the case that the conditions participants trained were learned differently, meaning that one different learning rate $\alpha(a_j)$ or $\beta(a_j)$, and not α and β existed per condition. Besides this, we also expected the necessity of having both Hebbian and non-Hebbian components to describe subject's habit learning, since they explain learning processes that occur in different stages of the task (cf. section 3.5).

Fourthly, we predicted the Pavlovian bias to be significantly different than zero which implies that, in average, participants had a bias towards pictures presented and had to overcome these biases through controlled actions engaging the brain regions referred in section 3.5.

Finally, if the Pavlovian bias does not present significance, we expect the use of the moving average method to reduce the noise of undesired processes and achieve significance on this parameter. Besides we also expect to verify the other first three hypotheses after applying the moving average method.

Thus, to perform a complete study of these hypotheses several computational models were developed (cf. table 3.2), and provided the necessary framework to use in the analysis of the behavioral data of the training version of the AAT.

Number	Model	Likelihood function
1	Model assuming that the average value of the reaction times did not change during training	Normal
2	Model assuming that the average value of the reaction times did not change during training	Log-Normal[9]
3	Model with two learning rates for both conditions	Normal
4	Model with two learning rates for both conditions	Log-Normal
5	Model with one learning rate (α) per condition	Normal
6	Model with one learning rate (α) per condition	Log-Normal
7	Model with one learning rate (β) per condition	Normal
8	Model with one learning rate (β) per condition	Log-Normal

Table 3.2: Number assigned to the different designed models where the maximum likelihood estimation was performed using different likelihood function centered at the predicted RTs [Normal: $RT_i \sim Normal(\widehat{RT_i}, \sigma^2)$; Log-Normal: $RT_i \sim LogNormal(Log(\widehat{RT_i}), \sigma^2)$].

3.6. Parameter estimation

A behavioral computational model is described by a set of parameters and equations which determine how the internal variables are updated on a trial-by-trial basis. Therefore the major objective, when using computational models to fit behavioral data, in our specific case, was to obtain the set of values of the model's parameters which better explained the data from each subject [106].

Thus, we focused our attention on computing the probability of a given set of parameters θ_s, given the individual data D_s and the model M, *i.e.*, $P(\theta_s|D_s, M)$, also known as the posterior probability. More concretely, to estimate the optimal parameters, we wanted to maximize this quantity.

According to Bayes' rule, the posterior probability is given by (equation 3.10).

$$P(\theta_s|D_s, M) = \frac{P(D_s|\theta_s, M) \cdot P(\theta_s|M)}{P(D_s|M)} \tag{3.10}$$

Since the model evidence ($P(D_s|M)$) is a normalization constant, it can be ignored during the process of estimating the optimal parameters. The product between the likelihood of the predicted data and the prior probability, on the other hand is essential. While the first is easily computed, the second is not, and in fact we did disregard this term when optimizing the parameters. This seemed to be the optimal solution because the prior was treated as non-informative, since, in our case, there was not reliable *a priori* information regarding this quantity [107].

Thereby maximum likelihood estimation was performed instead of a maximum a posteriori estimation (equation 3.11).

$$\widehat{\theta_s} = argmax_{\theta_s} P(D_s|\theta_s, M) \tag{3.11}$$

The result from $P(D_s|\theta_s, M)$ is normally extremely small. For this reason, it is numerically more stable and computationally less heavy to compute the logarithm of this quantity. Consequently, parameter estimation was performed through the maximization of the Log-likelihood (LLH).

Since this estimation method presents a high computational load, we initially assumed that the observed RTs followed a Normal distribution centered at the predicted RTs and tried to estimate the

[9] This likelihood function was chosen according to the results presented throughout the sections 4.1 and 4.2.

parameters using the Ordinary Least Squares (OLS) method, which estimates the best set of parameters through minimization of the residual sum of squares (SSE), (cf. equation 3.12).

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.12}$$

Where $\boldsymbol{y}$ represents the vector of the observations of our response variable and $\hat{\boldsymbol{y}}$ represents the vector of predicted values of $\boldsymbol{y}$.

However, despite very common [37] [108] [109], we soon realized that such a method would not be suitable for our case, since the RTs did not seem to follow a Normal distribution [89].

Considering the results of the exploratory data analysis which we performed (cf. section 4.1), our real concern regarded the homogeneity of the variance of the errors. This was of specific interest, because we needed a method that estimated the parameters considering that the distribution of the RTs would not have symmetrical tails and, in our specific case, would present a heavy right tail.

Therefore we resorted to the maximum likelihood estimation method, by using two different likelihood functions [Normal: $RT_i \sim Normal(\widehat{RT}_i, \sigma^2)$ and Log-Normal: $RT_i \sim LogNormal(Log(\widehat{RT}_i), \sigma^2)$], in order to confirm the results of the exploratory data analysis which we performed. These indicated that using a distribution with a heavy right tail would provide significantly better results (cf. section 4.1).

It is also important to refer that the correct estimation of the parameters and respective likelihood of the predicted data was dependent on the fixed variance of the distribution which we imposed to each subject. Consequently, in an initial phase, we used arbitrary variances for both likelihood functions and then, in a second phase, making use of the predicted data we estimated the variance which maximized the likelihood for both distributions (described in sub-section 3.5.1) in order to then accurately compare the results obtained through different computational models.

3.6.1. Variance: punctual estimation

The variance (σ^2) is a very important measure when performing regressions, in general, since its estimation indicates the variability of the response variable. Furthermore, as explained in the previous section, in our computational models, this measure was essential for the fitting of the empirical RTs through maximum likelihood estimation.

In order to estimate σ^2 of each observation (RT_i), we first have to convert equation 3.7 into a regression model, obtaining equation 3.13.

$$RT_i = (\Delta w_i)^{-D} + E + \varepsilon_i \tag{3.13}$$

Then, we need to compute the sum of squared deviations. However, it is important to notice that each RT_i comes from analogous distributions with different means ($\widehat{RT}_i$) which depend on Δw_i. Therefore, the residuals (estimates of the errors ε_i) are calculated through equation 3.14 [110].

$$RT_i - \widehat{RT}_i = e_i \tag{3.14}$$

Thus, the residual sum of squares (SSE) is given by (equation 3.15):

$$SSE = \sum_{i=1}^{n}(RT_i - \widehat{RT}_i)^2. \tag{3.15}$$

Consequently, the residual mean square (MSE) is obtained by (equation 3.16):

$$\widehat{\sigma^2} = MSE = \frac{\sum_{i=1}^{n}(RT_i - \widehat{RT_i})^2}{n-p}. \quad (3.16)$$

Where p is the number of parameters to be estimated in the model that computes the $\widehat{RT_i}$s.

However, this previous reasoning only applies for normally distributed errors [110]. In our specific case, we assumed that the RT_i could also follow a Log-Normal distribution as explained and demonstrated in section 4.1. Therefore, to correctly use the maximum likelihood method we had to find an estimator for the variance of the predicted RTs.

Considering equations 3.17 (the estimated variance of a normally distributed sample), 3.18 and 3.19 (the estimated variance and mean of a log-normally distributed sample, respectively [111]).

$$\widehat{S^2} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n} \quad (3.17)$$

$$\widehat{S^2} = \frac{\sum_{i=1}^{n}(Log(Y_i) - \hat{\mu})^2}{n} \quad (3.18)$$

$$\hat{\mu} = \frac{\sum_{i=1}^{n} Log(Y_i)}{n} \quad (3.19)$$

From the analogy found between equation 3.17 and 3.18 and the rationale explained to obtain equation 3.16, our estimator was derived. Therefore, in this case we estimated the σ^2 through equation 3.20:

$$\widehat{\sigma^2} = MSE = \frac{\sum_{i=1}^{n}[Log(RT_i) - Log(\widehat{RT_i})]^2}{n-p}. \quad (3.20)$$

However, we are aware that this might be a sub-optimal estimator, since we did not prove that it was an unbiased estimator, the rationale made is not devoid of empirical validation.

3.6.2. Non-linear optimization

Since we had to use a non-linear function to fit the participant's behavioral data, we resorted to a non-linear optimization routine to proceed. Thereby in an initial phase the *fmincon* function of MATLAB was selected.

This routine tries to minimize the value of a specific function, while satisfying determined constraints regarding the set of parameters to be optimized. Therefore, we do not only have to provide the set of parameters of the model to the optimization routine, as also the initial conditions of the parameters at each iteration of the optimization routine, an objective function describing the model (cf. section 3.4), and the boundaries delimiting parameter space.

Regarding the estimation method to be applied, we initially resorted to the OLS to optimize the subject-specific parameters. However, due to the reasoning explained in section 3.6, we ended up using the *fmincon* function to perform maximum likelihood estimation.

In each iteration, we made the function either report the value of a local minimum found "near" the initial conditions and the respective set of parameters that originated that minimum, because it had achieved a pre-specified requirement, or the lowest value found until *fmincon* had reached the termination criteria. Thus, there is no guarantee that the global minimum is found through this routine for an objective function like ours, since this method is not the most accurate one in finding minima in non-smooth surfaces. In order to try to overcome this short-coming and to enhance the probability of

finding the global minimum, we used a large number of iterations with different starting points widely dispersed and also modified some of the default stopping criteria, and verified that the best reported minimum had been found in an iteration much lower than the number of iterations ran for each subject.

Notwithstanding, given the fact that this optimization was very time-consuming, we decided also to implement the model and respective routines in RSTAN (cf. sub-section 3.5.2).

The major advantage of using this imperative language is that, like C or Fortran, it is based on assignment, loops, conditionals, local variables, object-level function application and array-like data structures. Thus, although it was much more difficult to implement higher-order functions using this type of language, in what concerns fastness and efficacy it was much better than MATLAB, since the objective was to find a global minimum in a way similar to the one explained above [105].

Therefore, we used one of the Stan's optimizers, more specifically the Limited memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm, which, briefly, is one method that uses an estimation to the inverse Hessian matrix to steer its search through variable space and is also considered to be particularly well suited for optimization problems with a large number of variables, which was our case.

Moreover, we verified that the same results were obtained for test subjects, whose data were fitted using both methods. For all these reasons we decided to use this routine in the final analysis of this thesis. For more detailed description of this method see [105].

3.7. Model comparison

Posterior to the estimation of the parameters it is imperative to validate the created models. Additionally, it is important to state that different methods were used for the mixed and the novel behavioral models developed.

3.7.1. Mixed models

According to the Neymann-Pearson lemma, the most informative statistic to compare models is the ratio of the probabilities of different models given the observed data [112].

This probability, also known as model evidence ($P(D_s|M)$), is consensually the most robust and reliable measure for model comparison [106]. It is calculated according to equation 3.21.

$$P(D_s|M) = \int P(D_s|\theta_s, M) \cdot P(\theta_s|M)\, d\theta_s \tag{3.21}$$

As can be noted, the model evidence does not depend explicitly on the parameters because the likelihood of the data is averaged out according to their prior probability. Moreover, $P(D_s|M)$ is a probability distribution, and so, when integrated over all datasets, it must be equal to 1. Thus, more complex models, that usually present a better fit to more datasets, will typically assign lower evidence to each dataset, causing the measure to favor simpler models in the absence of strong contradictory information. Consequently, the use of the model evidence to perform model comparison does not lead to overfitting problems, as it would occur if the maximum likelihood, for instance, was used instead [106] [112].

However, usually, equation 3.21 is analytically intractable and numerically difficult to compute for the models which we are interested in comparing. In addition, as we already mentioned, usually there is no

valid information regarding the prior probabilities of the parameters, making this computation impossible or, at least, extremely unreliable in case the assigned priors do not suit the data [112].

Having this in mind we tried to overcome these difficulties by using approximations to model evidence. These are presented next.

The Akaike's Information Criterion (AIC) which is defined by equation 3.22:

$$AIC = -2 \cdot (Log\big(P(D_s|\theta_s, M)\big) - k). \tag{3.22}$$

The Bayesian Information Criterion (BIC) computed according to equation 3.23:

$$BIC = -2 \cdot (Log\big(P(D_s|\theta_s, M)\big) - \frac{k}{2} \cdot Log\, n). \tag{3.23}$$

Where k represents in both equations the number of parameters and n represents the number of observations.

As can be noticed, both criteria present two main quantities, an accuracy-related term (which makes use of the maximum likelihood estimate), and one related to the complexity of the model (which objective is to penalize the model's complexity) [113] [114].

If we take each criterion individually, we might not reach consensus regarding model selection. However, when both are simultaneously used, we can extract important information given the fact that, in general, the AIC penalizes the complexity of the model not sufficiently enough and the BIC penalizes it too much [113] [114].

Having the maximum likelihood estimates, evaluating the goodness of the fit of distinct mixed models through the analysis of their model evidence approximations was consequently straightforward.

3.7.2. Behavioral models

Here, the first question could be if the models should be treated as fixed effects (meaning one model formulation would be enough to characterize an entire group) or random effects, in which each subject could be characterized by a specific model and distribution.

Since one of the objectives of this thesis was to capture subjects' behavior during the training task and the fixed effects approach was used through the analysis of the mixed models, we decided to treat the models as random effects. Therefore we used a hierarchical model to assess their probability densities.

The hierarchical model considered each subject (s) to be described by a set of binary variables (m_{sk}), which in turn assigned the model (k) to that subject. In the model, those variables are generated by a multinomial distribution with parameters r (the model probabilities to be estimated). This hierarchical model also states that the model probabilities follow a Dirichlet distribution with parameters α, which are associated to unobserved occurrences of the models in the population [115].

Then to compute the effective number of subjects for whom a specific model generated the data, we had to invert the hierarchical model to obtain the values of α and subtract the prior. To perform this inversion, a variational Bayes algorithm, which is described in [112] was used.

Having α, the estimation of the variables of interest is straightforward and the model selection procedure is concluded (figure 3.7). The most important thing here was, that for performing all these

procedures, we only had to estimate the model evidences ($P(D_s|M_k)$), which were computed using the approximation of the BIC criterion (cf. sub-section 3.7.1).

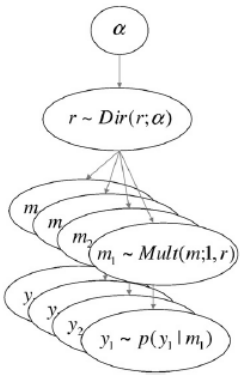


Figure 3.7: Hierarchical Bayesian model with random effects used to perform Bayesian model selection [115].

Thus, using the α values, the expected values of the probabilities of the models (r_k), and the exceedance probabilities (EPs or Φ_k) are easily computed. These are a quantification of the confidence that a particular model is more likely than the remaining, given the observed data (y) (equation 3.24, where K is the total number of models).

$$\Phi_k = P(r_k > r_{j\neq k}|y), \forall j,k \leq K \tag{3.24}$$

However, these quantities are still not sufficiently robust because their computation is not protected against the assumption that the observed differences in model frequencies may be caused by chance, which could be a plausible explanation [115]. Consequently, we could not use these quantities directly to perform the Bayesian model comparison (BMC), but used a quantity that protected the EPs against these disturbances.

This problem was raised by Rigoux *et al.* (2014) who made use of the concept of Bayesian omnibus risk (BOR) to obtain such quantities. This is conveyed into a value that measures the statistical risk of performing BMCs, directly quantifying the probability that the frequencies of the models were all the same and simply seemed to be different by chance [115]. This quantity was then used to obtain the protected exceedance probabilities (PEPs), through a Bayesian model average of the EPs (equation 3.25).

$$\widetilde{\Phi_k} = \Phi_k \cdot (1 - BOR) + \frac{1}{K} BOR \tag{3.25}$$

Thus, these quantities provided the essential information to properly perform the Bayesian model selection (BMS). An important remark regarding equation 3.25 is that if BOR tends to zero, it means that the hierarchical model of figure 3.7 is reliably better than chance. Another consequence is that if that is the case, the PEPs will be very similar to the EPs.

This method was performed in MATLAB, using the toolbox described in [116], which was modified to calculate the PEPs of the analyzed models. Besides that, the output of the *VBA_groupBMC* function was also modified, in order to more easily obtain this extra information and generate the plots of interest [49].

4

Results and Discussion

Contents

In this chapter, the behavioral data collected during the performance of the three versions of the AAT is analyzed.

4.1. Exploratory data analysis

According to Ratcliff (1993), depending on the design of the task, the RTs can follow different distributions [86]. Consequently, all subsequent analyses should be done taking into account the hints provided by the results presented in this section.

During our exploratory data analysis, first we concatenated the RTs of the different sessions for each subject. Then, the function *histDist* and next the *Kolmogorov-Smirnov* test were applied to each subject as described in section 3.3. *Kolmogorov-Smirnov* tests were performed to test the null hypothesis that the RTs followed a Normal distribution or a Log-Normal distribution. The $p - value$ obtained, in general, were highly significant, meaning that the null hypotheses should be rejected (cf. appendix A3).

Figure 4.1 depicts the outcome of using the function *histDist* on 8th subject of the negative group.

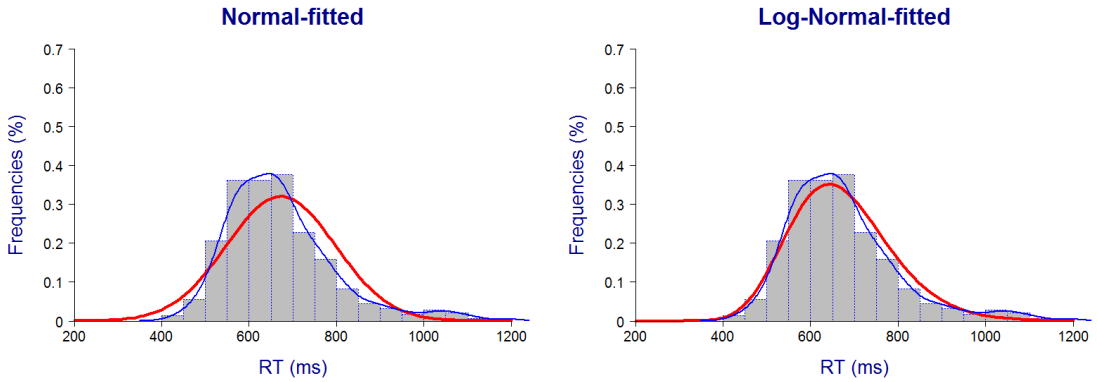


Figure 4.1: Representation of the output provided by function *histDist*, on the behavioral data of the 8th subject of the negative group. The blue line is an interpolation performed to the histogram and the red line in (a) represents the fitted Normal distribution whose parameters (μ and σ) were estimated from the RTs and (b) represents the fitted Log-Normal distribution whose parameters (μ and σ) were estimated from the RTs.

The results from the *Kolmogorov-Smirnov* test for the 8th subject of the negative group corroborated what was above mentioned (Normal fit: $p - value < 0.001$; Log-Normal fit: $p - value = 0.017$).

Figure 4.2 depicts the outcome of using the function *histDist* on the 8th subject of the positive group

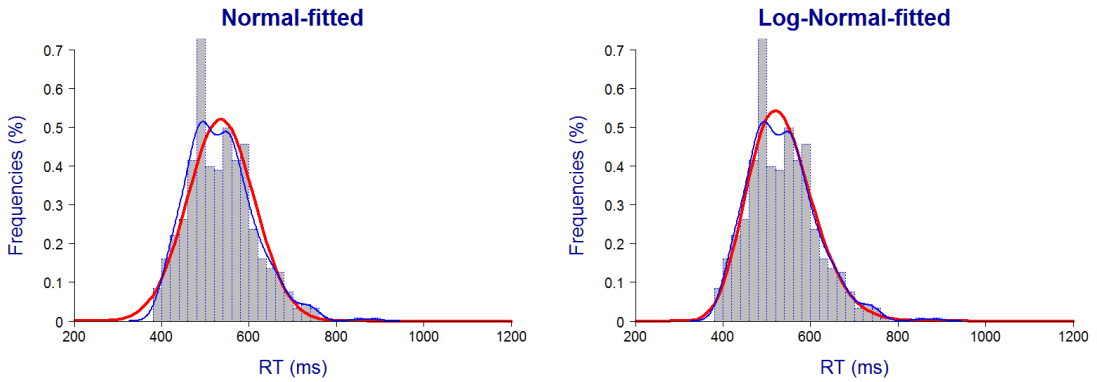


Figure 4.2: Representation of the output provided by function *histDist*, on the behavioral data of the 8th subject of the positive group. The blue line is an interpolation performed to the histogram and the red line in (a) represents the fitted Normal distribution whose parameters (μ and σ) were estimated from the RTs and (b) represents the fitted Log-Normal distribution whose parameters (μ and σ) were estimated from the RTs.

Again the results from the *Kolmogorov-Smirnov* test for the 8th subject of the positive group corroborated what was above mentioned (Normal fit: $p - value = 0.038$; Log-Normal fit: $p - value = 0.045$), despite the fact they were not highly significant.

These results actually were in line with our supposition that the RTs would not follow a Normal distribution, but might follow a Log-Normal one. This should be the case because the sum of Log-Normal distributions did not result in a Log-Normal distribution unless these distributions were identically independent so that it an approximation could be applied [117]. Actually we did not assume that the different sessions would be independent (cf. sub-section 4.2.3).

Then, we used the same procedure above for each session of all participants (cf. appendix A3). Figure 4.3 and figure 4.4 depict the outcome of applying the function *histDist* to each session of the 8th subject of the positive group.

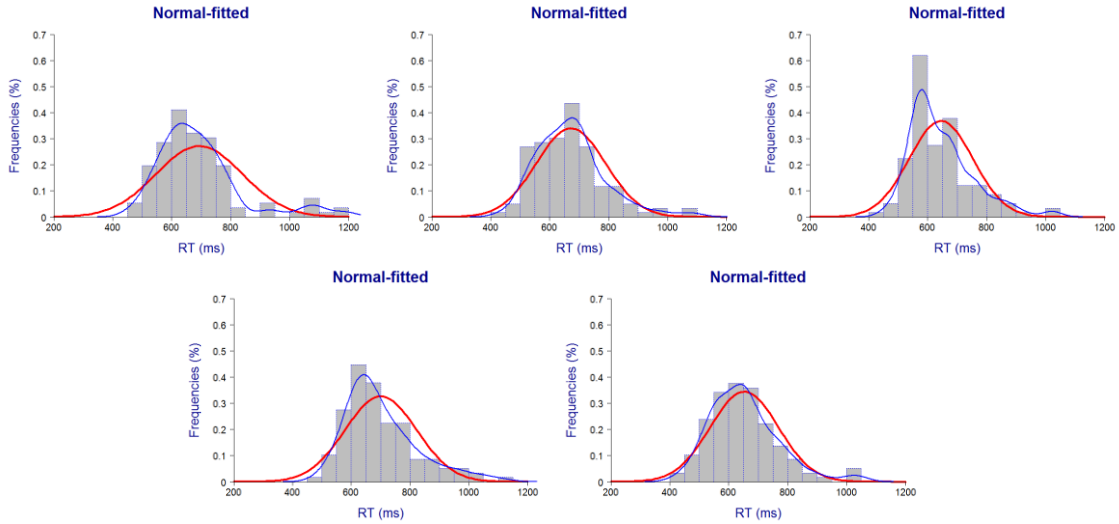


Figure 4.3: Representation of the output provided by function *histDist* on the behavioral data of the 8th subject of the negative group. The blue line is an interpolation performed to the histogram and the red line represents the fitted Normal distribution whose parameters (μ and σ) were estimated from the RTs in (a) the first session, (b) the second session, (c) the third session, (d) the fourth session and (e) the fifth session.

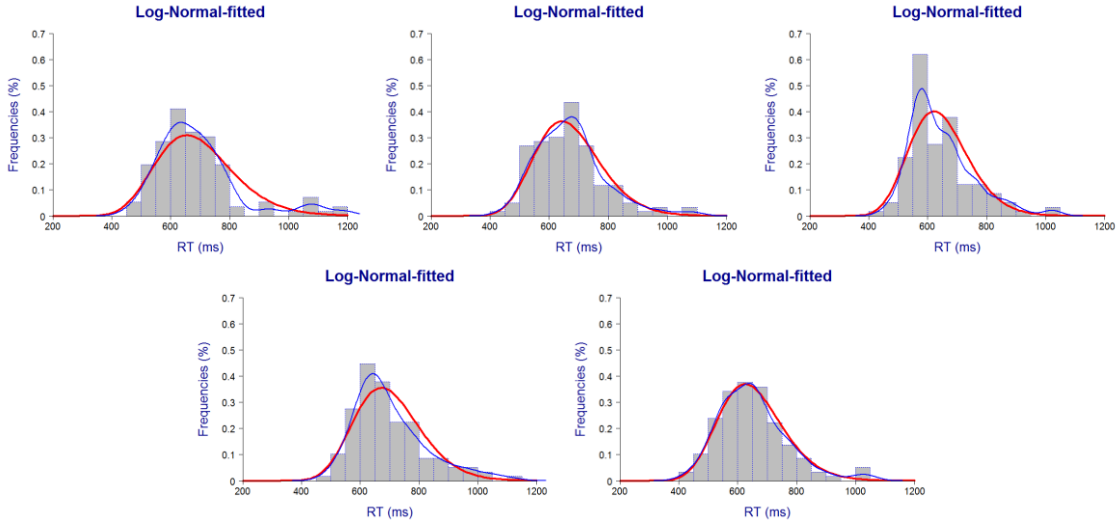


Figure 4.4: Representation of the output provided by function *histDist* on the behavioral data of the 8th subject of the negative group. The blue line is an interpolation performed to the histogram and the red line represents the fitted Log-Normal distribution whose parameters (μ and σ) were estimated from the RTs in (a) the first session, (b) the second session, (c) the third session, (d) the fourth session and (e) the fifth session.

In table 4.1 we present the results of the *Kolmogorov-Smirnov* test for this subject.

Subject	Group	Distribution	$p-value$ of the Kolmogorov-Smirnov test				
			Day 1	Day 2	Day 3	Day 4	Day 5
Subject 8	Negative	Normal	0.027	0.109	0.033	0.062	0.145
		Log-Normal	0.285	0.456	0.087	0.216	0.595

Table 4.1: Results of the *Kolmogorov-Smirnov* test applied to each session (day) of the 8th subject of the negative group.

The results of table 4.1 are not conclusively regarding the distribution which the RTs followed. Nonetheless, by analyzing the individual results of all participants (cf. appendix A3) more carefully, we inferred that the RTs would be more prompted to be log-normally[10] distributed than normally distributed. This gave an important hint and alerted us to neither use the OLS, nor the normal probability density function during the optimization of the computational models' parameters via maximum likelihood estimation (cf. section 3.6).

In the following sections, several of the obtained results will support this analysis.

4.2. Model-free analysis

In this section, we will describe the model free analysis of the behavioral data of the second sample. The behavioral data and respective analysis are divided according to the version of the AAT from which the data was acquired. It is important to mention that the behavioral data acquired from the different versions of the AAT, from the ratings and from the practical test were analyzed by the mixed model approach. Moreover, throughout the analysis we distinguish two distinct groups: the negative group (which corresponds to the participants trained to approach negative pictures and to avoid neutral pictures) and the positive group (which corresponds to the participants that trained to avoid positive pictures and to approach neutral pictures).

The analyses of this section were performed in the software for statistical computing R (R version 3.1.1).

For testing the relevance of the parameters estimated by the mixed models, we make use of the *anova* function from the package *lmer* which performs the usual analysis of variances, but takes into account that we are using mixed models.

Consequently, it makes use of the Satterthwaite approximation for degrees of freedom, which estimates an "effective degree of freedom" for a probability distribution formed from several independent normal distributions where only estimates of the variance are known.

For all the comparisons and contrasts performed during the analysis, the *glht* function from the package *multcomp* was used. This provides the *z-score* of the test and the respective $p-value$.

As a final note we wanted to say that one expression of the form $A * B$ represents $A + B + A \times B$, *i.e.*, it includes both the main effects and the interaction between the two variables. Besides this, the random effects in the mixed models will be displayed within parentheses.

[10] Although we only present results for the Log-Normal and Normal distributions, other distributions with a right heavy tail (such as the ex-Gaussian and the Inverse Gaussian) were tested. However, the best results were not significant. For that reason we did not present them.

4.2.1. Arrow version of the AAT

Our first concern was to check whether there was any significant action bias *a priori* at group level. The RTs acquired from the arrow version of the task were analyzed through the implementation of a mixed model, with the following design (equation 4.1):

$$rt = action * group + (1|subject). \quad (4.1)$$

It is important to state that the RTs used for the analysis were submitted to the pre-processing mentioned in section 3.3.

Given this, and before presenting the results, let us formally present our hypotheses:

- We did not expect a major difference between the two actions. Additionally, since we had a balanced sample, we did noeither expect any major difference between groups nor a difference between groups concerning the effects of movement.

The analysis of the mixed model designed above showed that there was neither evidence for a main effect of action [$F_{(1,661.11)} = 1.112, p - value = 0.292$] nor a main effect of group [$F_{(1,35.83)} = 0.001, p - value = 0.970$]. The interaction term showed also non-significance [$F_{(1,661.11)} = 3.502, p - value = 0.062$]. These results indicated that there was no existent task-related action bias. So, we assumed that all differences that we saw in the training and assessment versions of the AAT are due to the used stimuli and the psychological and cognitive processes involved when performing the task.

Figure 4.5a depicts the mean RTs in ms, between the different actions, within each group, and figure 4.5b depicts the ratio between the reaction times associated with avoidance and approach reactions.

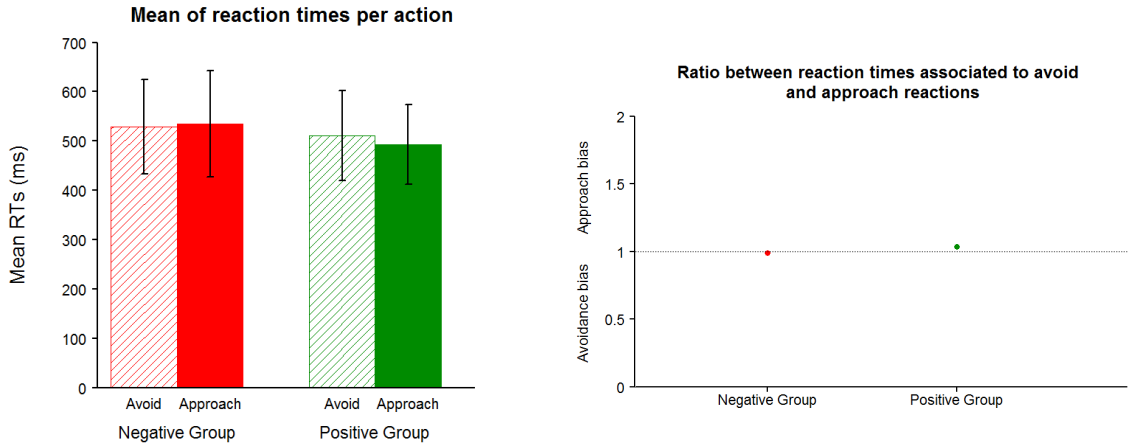


Figure 4.5: Graphical results of the analysis of the arrow version of the AAT. (a) Depicts the mean RTs in ms, for the different actions within each group, and (b) depicts the ratio between the reaction times associated with avoidance and approach reactions.

Although there is a small group difference observable, there was no major difference between actions within the group as reported above. Moreover, as it was shown in Figure 4.5 (b), that the ratio between the RTs of the two actions was extremely close to 1.

Regarding the individual level, a visual overview is presented in the appendices (cf. appendix A4). Although there is reasonable difference between the two actions in some subjects, the results above show that this individual difference is dispersed. Given this fact no correction was done in the RTs obtained in the assessment version of the task.

4.2.2. Assessment version of the AAT

Next we present and discuss the results regarding the assessment version. The data used was pre-processed according to section 3.3.

Here, we will try to anticipate the effects of the training and verify if there is evidence of effects due to the training performed during 5 consecutive days. Therefore, our hypotheses are the following:

- Before initiating the training, participants will have an approach bias for positive pictures, an avoidance bias for negative ones and no bias for the neutral ones. Additionally, these trends should not change for the pictures of categories that participants do not train.
- After the training, participants of the negative group should have an approach bias for negative pictures and an avoidance bias for the neutral ones, while the participants of the positive group should have an approach bias for the neutral pictures and an avoidance bias for the positive ones.
- We also wanted to verify whether, after the training, participants were able to generalize the learning to similar pictures, *i.e.*, we wanted to verify if there was a generalization effect in behalf of the trained condition. More concretely we expected for instance that participants of the negative group, when presented with negative pictures similar to the ones they trained, showed a similar (even though weaker) behavior when compared to the trained pictures. The same rationale is applied to the other picture categories.

In order to test for these hypotheses the following mixed model was design (equation 4.2):

$$bias = category * session * trained + generalized + (1|subject). \quad (4.2)$$

Regarding the variables used:

- $bias$ represents the difference between the RTs assigned to the avoid action and the RTs assigned to the approach action. A positive value indicates an approach bias, a negative one indicates an avoidance bias.
- $category$ is a categorical variable with four levels: 1 for positive pictures, 2 for negative pictures, 3 for neutral pictures used in the positive training group and 4 for the neutral pictures used in the negative training group.
- $session$ is a categorical variable with two levels: 1 for the assessment performed before training and 2 for the assessment performed after training.
- $trained$ is a categorical variable with two levels: 0 for the trials with untrained pictures and 1 for the trials with trained pictures.
- $generalized$ is a categorical variable with two levels: 0 for the trials with pictures that were not used for generalization, 1 for the trials with pictures used for generalization.

Before analyzing this model, it is important to refer that the significance level was set to a $p - value < 0.050$ and no correction for multiple testing was performed due to the explorative character of the analysis.

The analysis of the mixed model designed did not show evidence for main effects of category [$F_{(3,415.78)} = 1.790, p - value = 0.150$], nor of session [$F_{(1,394.85)} = 0.120, p - value = 0.730$], nor of trained vs. untrained pictures [$F_{(1,394.82)} = 0.100, p - value = 0.750$], nor of generalization [$F_{(1,394.93)} =$

$1.490,\ p-value=0.220$]. These results are not surprising, since the bias should not be predicted by these main effects *per se*. In what concerns the 2 way interactions, again, there were no significant effects (category by session [$F_{(3,394.85)}=0.570, p-value=0.640$], category by trained [$F_{(1,415.9)}=1.270, p-value=0.280$], nor session by trained [$F_{(1,394.85)}=0.510, p-value=0.470$]), which in fact are good results: If any of these interactions was significant, it would mean that the third variable, that is not present, was not necessary to describe the respective findings.

The 3 way interaction, the term of utmost importance, also did not show evidence of being significant [$F_{(3,394.85)}=1.300, p-value=0.270$]. This interaction has a high relevance because it contains information regarding whether or not from one session to another, in a specific condition, there are differences between trained and untrained pictures. However, the significance of this interaction strongly depends on the best possible reduction of noise. Since our results indicated that the task design did not eliminate noise, *i.e.,* uncontrolled influences, in a sufficient manner (see figure 4.6), we continued with the post-hoc analysis despite the non-significant 3 way interaction.

Figure 4.6 depicts the reaction biases before and after the training of the positive group, towards negative and neutral pictures they did not train, and for the negative group, towards the positive and neutral pictures they did not train (left side), and reaction biases before and after the training of positive group, towards the conditions they trained, and for the negative group, towards the conditions they trained (right side).

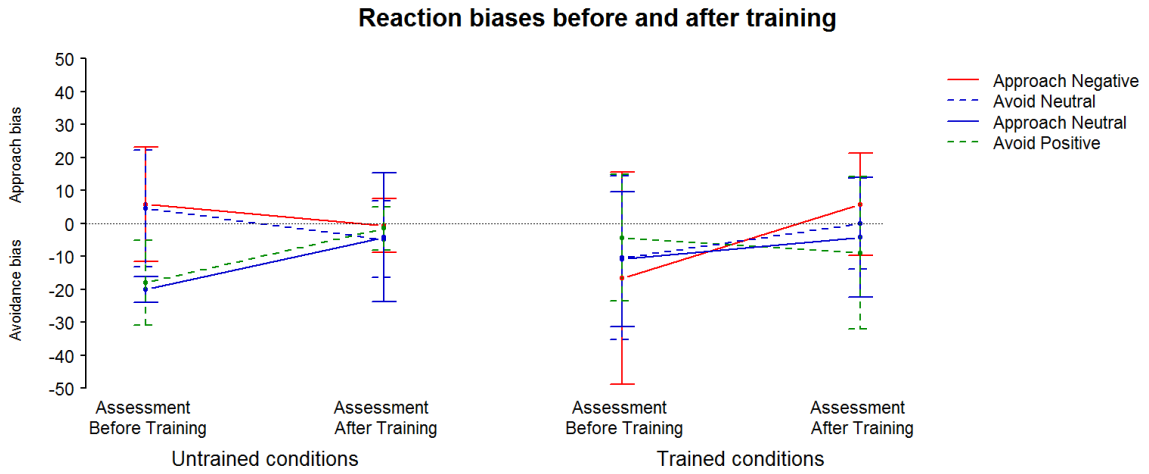


Figure 4.6: Results from the assessment version of the AAT that depict the reaction biases before and after the training of the positive group, towards negative and neutral pictures they did not train, and for the negative group, towards the positive and neutral pictures they did not train (left side), and reaction biases before and after the training of positive group, towards the conditions they trained, and for the negative group, towards the conditions they trained (right side).

However, visual inspection of figure 4.6 shows that it is possible to identify some of the tendencies enunciated in the hypotheses above. The before mentioned noise is also clearly visible in terms of the pronounced individual variability within groups, represented by the long error bars.

Therefore, contrast tests were performed. The first contrast was to check if, before the training, there was any approach bias for positive pictures in both groups. It was shown that this was not the case. Then, we checked for avoidance bias for the negative pictures. The results showed that none of the points were significantly different from zero ($p-value>0.200$).

Another result that we did not expect was to have biases towards the neutral pictures before training. Moreover, this is not only depicted in the figure above, but it also shows significance for a general avoidance bias [$z - value = -2.700,\ p - value = 0.007$], especially for the data point that represents the approach neutral condition in the left side of the figure above [$z - value = -2.960,\ p - value = 0.003$]. This result might be surprising given the effort we put into selecting the pictures. Therefore, this result will be discussed again when having a closer look at the analysis of the ratings (cf. sub-section 4.2.4).

Notwithstanding the above results, we decided to perform another plot (figure 4.7) in which it is depicted the differences in reaction biases before and after training of the positive group, towards negative and neutral pictures they did not train, and for the negative group, towards the positive and neutral pictures they did not train (left side), and the differences in reaction biases before and after training of positive group, towards the conditions they trained, and for the negative group, towards the conditions they trained (right side).

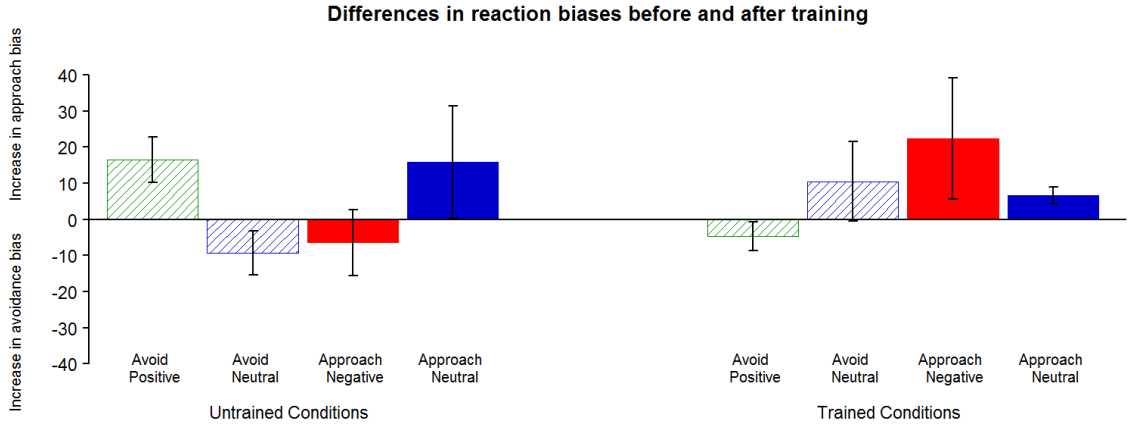


Figure 4.7: Results from the assessment version of the AAT. The differences in reaction biases before and after training of the positive group, towards negative and neutral pictures they did not train, and for the negative group, towards the positive and neutral pictures they did not train (left side), and the differences in reaction biases before and after training of positive group, towards the conditions they trained, and for the negative group, towards the conditions they trained (right side).

From figure 4.7 we infer that the majority of the tendencies seem to go towards the direction of what people trained. This, in fact, is an indicator that the training, might have changed the original reaction tendencies. Actually, figure 4.7 shows in a more straight-forward way whether the trends we hypothesized are present. Although the aforementioned 3 way interaction is not significant, the tendency between the red bars is a trend [$z - value = 1.720,\ p - value = 0.090$] and the tendency between the green ones also points to the desired difference, although it is not a trend [$z - value = -1.580,\ p - value = 0.120$]. It is important to refer that these results are better than the ones in figure 4.7, because by taking the difference between sessions, we apparently also subtracted the noise within subjects. Nevertheless, the noise between subjects is still present as it can be seen from the error bars.

Taking into account the note above on the reduction of the noise, another graphic (cf. figure 4.8) was performed in order to visually inspect what was said regarding the pictures used for generalization.

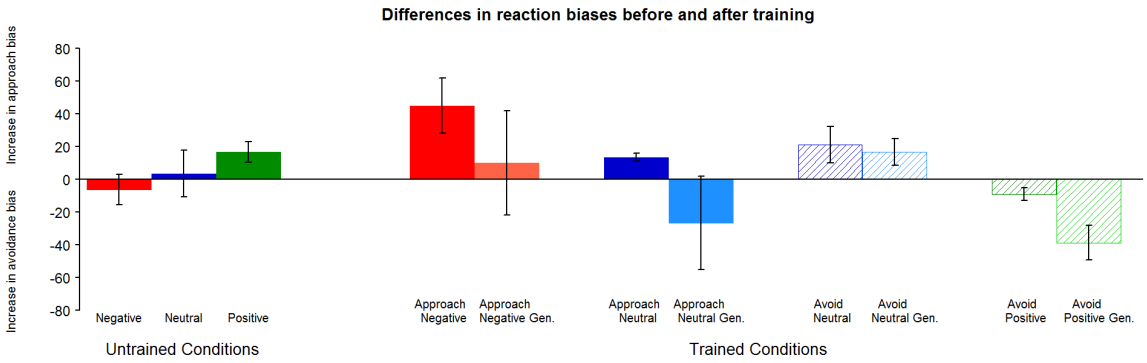


Figure 4.8: Results from the assessment version of the AAT. The differences in reaction biases before and after training of the positive group, towards negative and neutral pictures they did not train, and for the negative group, towards the positive and neutral pictures they did not train (the three initial bars) and the differences in reaction biases before and after training of positive group, towards the conditions they trained, and for the negative group, towards the conditions they trained, including results obtained for generalization, respectively (right side).

From figure 4.8 we could infer that the tendency towards generalization by the subjects might be present, at least in the case of the negative and positive pictures. In fact, by performing pairwise comparisons between pictures trained and pictures used for generalization, a trend was achieved in the negative case [$z - value = 1.890,\ p - value = 0.060$], while for the positive pictures this trend, event though not significant [$z - value = -1.540,\ p - value = 0.120$], was into the hypothesized direction. Besides this, we also tested, in trained conditions, for the difference between the first red and green bars. The results did not show significance [$z - value = -1.630,\ p - value = 0.104$], but provided a hint into the direction that the training worked and participants might have acquired new reaction tendencies. In fact, visual inspection lead us to presume that a bigger change was achieved for the negative group. For the neutral conditions, the above reported tests were also performed and the results did not yield significance ($p - value > 0.200$).

Finally, we also performed pair-wise comparisons between the values of the untrained conditions. The results showed that there is a significant difference between positive and negative [$z - value = 2.080,\ p - value = 0.037$] and between negative and neutral pictures [$z - value = 2.020,\ p - value = 0.043$]. These results indicate that there is an increase in the avoidance bias of participants of the positive group towards negative pictures, relative to an increase in the approach bias of participants of the negative group towards the positive pictures.

To sum up this quite extensive analysis, it is important to bear in mind for the following analysis that, even though most effects did not reach the desired level of significance, there is evidence for some training effects. In fact, given the high level of noise that is present and hints to instability of this implementation of the AAT (cf. section 5.1 for further discussion), we think that the results are showing reasonable trends. Nevertheless, we are fully aware that these results have to be seen with caution and only the execution of (the already planned) further studies will allow for any final conclusion.

4.2.3. Training version of the AAT

In order to proceed with the analysis of the behavioral data acquired, first, we had to decide how to treat the data of the different days. Two options were given: consider the dataset as coming from several

days that are independent from the others or assume that the datasets of different days were dependent and concatenate the datasets of different days, thus considering one big dataset for each subject.

According to the reasoning explained in previous sections (*e.g.*, section 4.1), we decided for the second option. This choice was also supported by findings provided by [28] [71] [118]. Furthermore all participants reported that they remembered the pictures they had trained in the day before and what they had to do, evidencing that besides habit learning there was also a strong evidence for episodic memory.

Then, as it was done in previous sections, after concatenation of the datasets of the 5 days, we processed the data, by removing the wrong trials and by excluding the outliers accordingly to the criteria specified in section 3.3.

Considering that this data was to be fitted by the computational model, first, we started to explore the data at group level, through the *nls* function of the *stats* package in R. To have then a visual overview, we resorted to *ggplot2* function. Figures 4.9, 4.10 and 4.11 depict the three different regressions we fitted the data with.

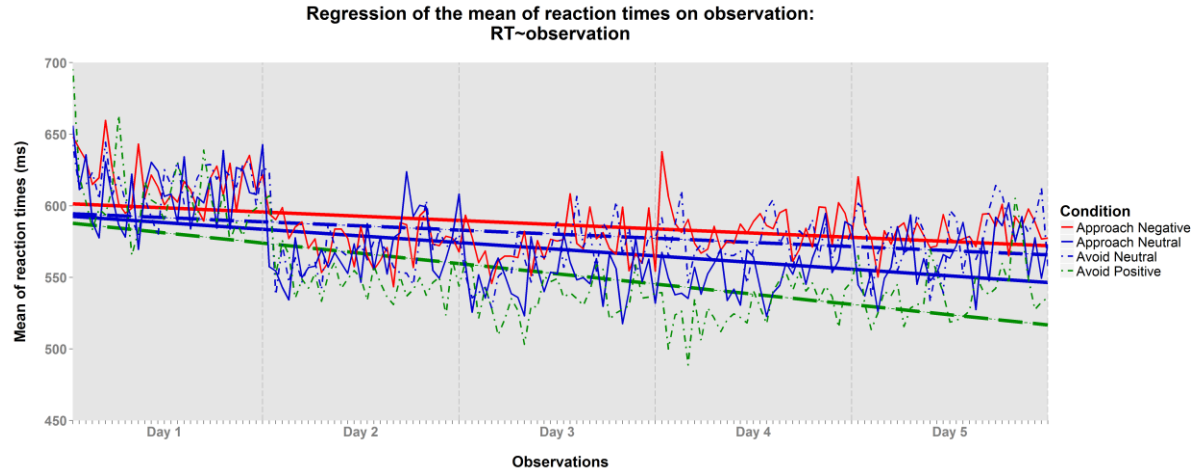


Figure 4.9: Results of the linear fit performed to each condition using the mean of the reaction times between pictures of that condition and between the subjects that trained that condition.

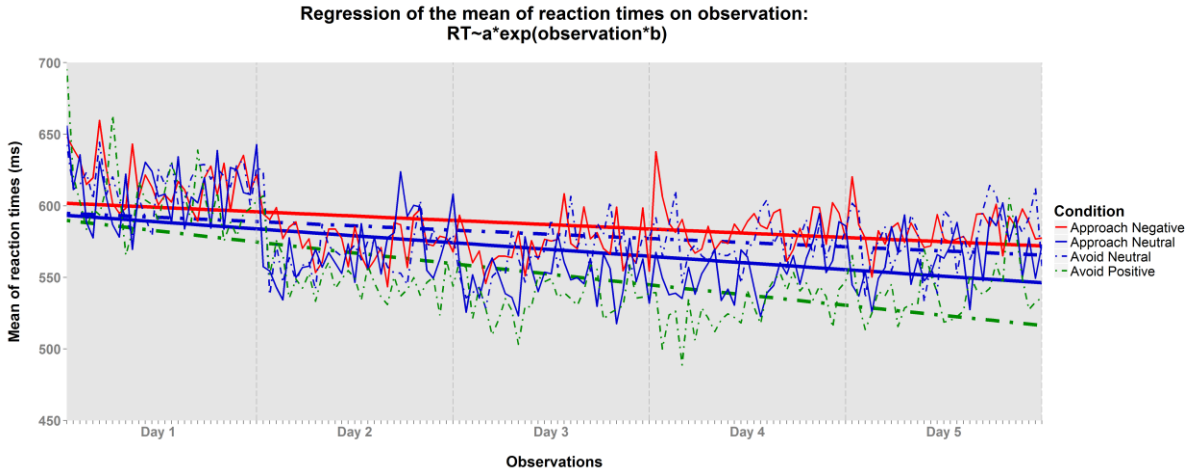


Figure 4.10: Results of the exponential fit performed to each condition using the mean of the reaction times between pictures of that condition and between the subjects that trained that condition.

The results of figure 4.10 were very similar to the results depicted by figure 4.9, because the parameter *b* had a value considerable lower than 1.

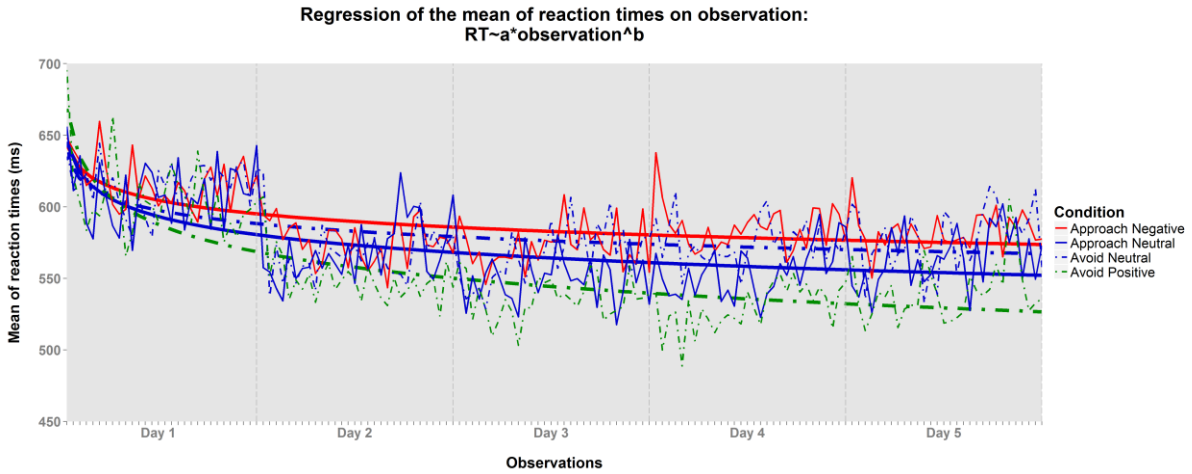


Figure 4.11: Results of the power law fit performed to each condition using the mean of the reaction times between pictures of that condition and between the subjects that trained that condition.

Notwithstanding, to understand the dynamics of the RTs and the influences of the subjects, we applied the mixed models approach to estimate parameters' significances. Thus, several models were fitted to the behavioral data in order to take into consideration the variability coming from the subjects as random effects and estimate the parameters of the group as fixed effects.

The predictors used in the models were:

- $trial$, a continuous variable that represents the trial number, whose maximum value is 600, since each session could have at most 120 trials and the participants were subjected to 5 sessions. Still, that value was never reached since the data was pre-processed and wrong trials and outliers were removed.
- $cond$, a categorical variable that represents the condition from which the RTs are coming. It has four levels *cond* = 1 corresponds to *approach negative* condition, *cond* = 2 corresponds to *avoid neutral* condition, *cond* = 3 corresponds to *avoid positive* condition and *cond* = 4 corresponds to *approach neutral* condition.

The different fits of the models are described in table 4.2. It is specified for each model the kind of trend fitted and the explanatory variables used to fit the data, as well as the kind of random effects that were tested, namely, two types of random intercepts and the random slope. For all models we provide the AIC and the BIC, beside the LLH.

Then, considering the results depicted in figures 4.9, 4.10 and 4.11, we established some test hypotheses:

- First, we expect to obtain better fits for the exponential and power law models when compared to the linear fits. This will be translated into better LLH values, which in turn provide evidence that the exponential and power law models will predict better the experimental data. Following this idea and since we were not interested in variable selection, we also predicted that the most complex models would be necessary to better explain the experimental data.

- Second, we expected the intercept parameter to be significant in each group and to be significant between the two conditions, *i.e.*, that there were two different intercepts within each group explained by the variability between conditions. Besides that, in the negative group, we predicted that the intercept for the negative condition would be higher than the one for the neutral condition. The same was expected for the positive group, but in this case the intercept for the positive condition would be higher than the intercept for the neutral condition.
- Third, we also expect the slope to be significantly negative in every model, meaning that the groups learned the trained conditions. Besides this, we hypothesize that there could be significant interactions between the slope and the condition, meaning that the variance in the slope would not only be explained by the difference between subjects, but also by the different conditions.

Model	Predictors	Random Intercept per condition	Random slope	AIC	BIC	Log-likelihood
Linear	$trial$	No*	No	252277.9	252309.7	-126134.9
		Yes	No	251980.7	252084.0	-125977.3
		Yes	Yes	251160.6	251470.6	-125541.3
	$trial + cond$	No*	No	252195.7	252251.3	-126090.8
		Yes	No	251977.3	252104.4	-125972.6
		Yes	Yes	251164.0	251330.9	-125561.0
	$trial * cond$	No*	No	252150.2	252229.7	-126065.1
		Yes	No	251931.2	252082.2	-125946.6
		Yes	Yes	251156.4	251514.0	-125533.2
Exponential	$trial$	No*	No	-16464.6	-16432.8	8236.3
		Yes	No	-16849.2	-16745.9	8437.6
		Yes	Yes	**-17712.1**	**-17569.1**	**8874.0**
	$trial + cond$	No*	No	-16576.3	-16520.7	8295.2
		Yes	No	-16853.1	-16725.9	8442.5
		Yes	Yes	**-17717.9**	**-17551.0**	**8880.0**
	$trial * cond$	No*	No	-16623.8	-16544.4	8321.9
		Yes	No	-16901.8	-16750.8	8469.9
		Yes	Yes	**-17727.6**	**-17536.8**	**8887.8**
Power Law	$trial$	No*	No	-16709.0	-16677.2	8358.5
		Yes	No	-17097.8	-16994.5	8561.9
		Yes	Yes	**-17853.9**	**-17710.8**	**8944.9**
	$trial + cond$	No*	No	-16822.0	-16766.4	8418.0
		Yes	No	-17101.7	-16974.6	8566.9
		Yes	Yes	**-17859.8**	**-17692.9**	**8950.9**
	$trial * cond$	No*	No	-16875.4	-16795.9	8447.7
		Yes	No	-17156	-17005	8597
		Yes	Yes	**-17870.8**	**-17680.1**	**8959.4**

Table 4.2: Model comparison between the 27 designed models we used to fit he behavioral data. These differed in model type (linear, exponential or power law), in number of predictors and in the random effects. The "No*" refers to the inclusion of just random intercepts.

From table 4.2, we could infer that the linear fits were the poorest. Therefore these models will not be discussed further.

Regarding the exponential and power law fits, it is important to refer that we used the logarithm transformation on the RTs. This was in fact a good approximation because by applying it, we obtained a new variable [$RT' = Log(RT)$] that would be normally distributed around the predicted RTs [119]. Actually the use of the function *lmer* required this. Therefore, we assumed this was a good approximation, since we also had support from the LLH values of this two models' fits.

Nonetheless, we were aware that this was an approximation and the following discussion was done with caution.

Another important remark regards the analysis of the AIC and BIC values, since in this case, we have to take into account that the lower they are the better the model is (cf. equations 3.22 and 3.23).

Focusing on both exponential and power law models' fits, the first thing that we conclude, when we keep the number of predictors, is that the better models require all the random effects, since they have better and congruent AIC, BIC and LLH values (highlighted values in table 4.2). However, a more detailed analysis of the highlighted values in table 4.2 indicated that, if we considered all the random effects, the AIC and BIC values were no longer congruent in the choice of the best model. Considering that the AIC can overestimate the goodness of model fit, while the BIC tends to penalize the model complexity more than the AIC (cf. sub-section 3.7.1), we observed that actually the AIC favors the most complex models, while the BIC tends to favor the simpler models.

Since these values are no longer congruent and considering that these values agreed that the better model was the one with all random effects, we decided to analyze the most complex model because we were interested in verifying the significance of all the predictors.

We analyzed the most complex models of the exponential and power law fits, because the developed computational model was based on a power law and both analyses were reported in prior studies; so, there was no clear preference for one of them.

Thus, the mixed models that we analyzed were respectively for the power law and the exponential fitting as follows (equations 4.3 and 4.4).

$$\mathrm{Log}(rt) = \mathrm{Log}(trial) * cond + (1 + \mathrm{Log}(trial) * cond|subject) \quad (4.3)$$

$$\mathrm{Log}(rt) = trial * cond + (1 + trial * cond|subject) \quad (4.4)$$

The analysis of equations 4.3 and 4.4 showed as expected that there are significant main effects for the exponential fit (trial: $F_{(1,20855.3)} = 324.99, p - value < 0.001$; cond: $F_{(3,103.1)} = 39.47, p - value < 0.001$) and for the power law fit (trial: $F_{(1,20820.7)} = 585.29, p - value < 0.001$; cond: $F_{(3,21.5)} = 4.12, p - value = 0.019$). Moreover, the interaction term also presented high significance for the exponential [$F_{(1,20855.1)} = 17.87, p - value < 0.001$] and for the power law [$F_{(1,20821.2)} = 20.13, p - value < 0.001$] fits.

Figures 4.12 and 4.13 depict respectively the exponential and power law models' fits to the behavioral data at group level.

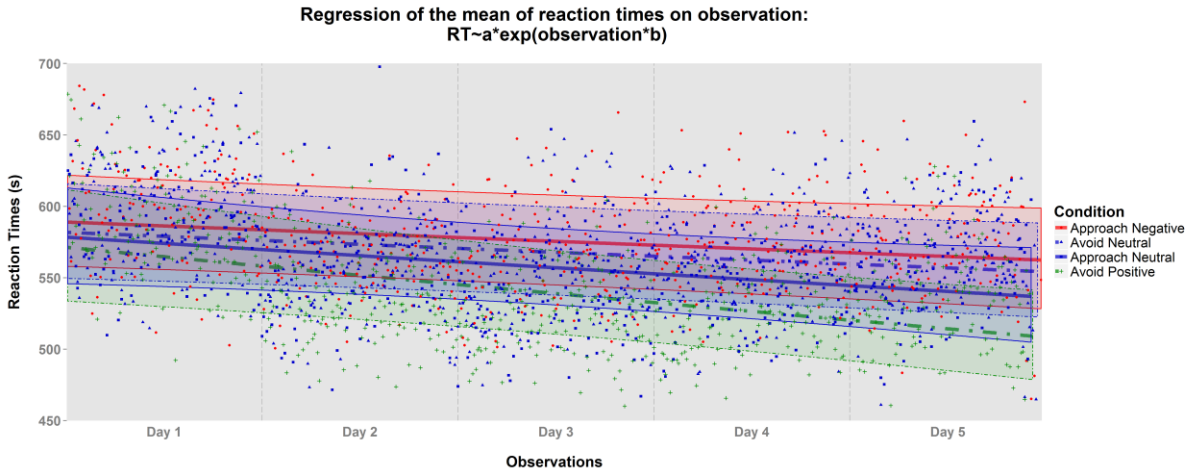


Figure 4.12: Results of the exponential model's fit performed to the behavioral data at group level.

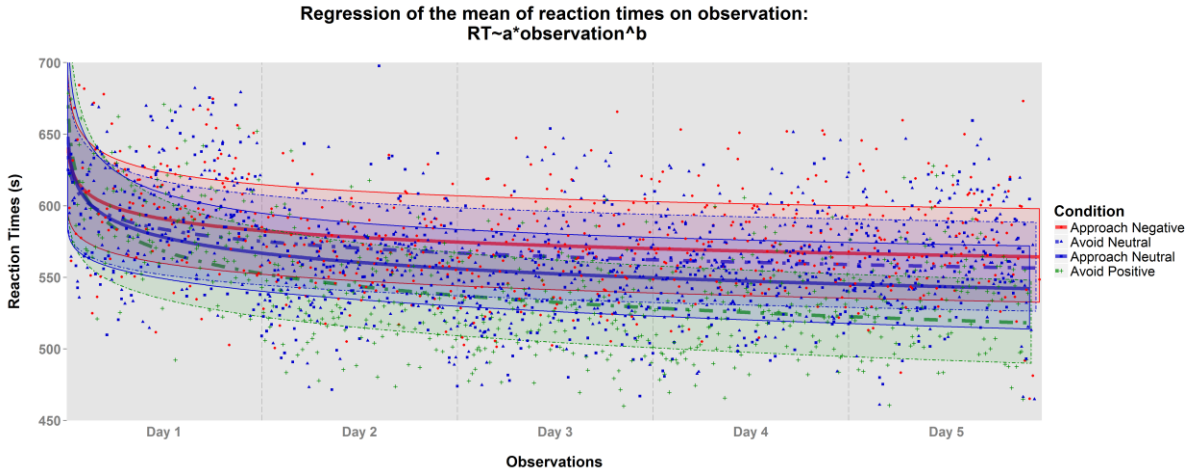


Figure 4.13: Results of the power law model's fit performed to the behavioral data at group level.

From figures 4.12 and 4.13 we confirmed that the highest starting point belonged to the negative condition, which was expected given the fact that this should be the most difficult condition to learn [120]. Nevertheless, posterior post-hoc tests showed evidence of a significant slope for negative condition for the power law fit [$z - value = -2.512,\ p - value = 0.012$] and for the exponential fit [$z - value = -2.200,\ p - value = 0.039$]. Another interesting fact is that the neutral conditions have very close starting points. When testing for this difference by means of post-hoc contrasts, the result shows, as expected, that they are not significantly different [$z - value = 0.543,\ p - value = 0.593$]. Following this rationale we also decided to test not only the remaining differences between intercepts, but also the differences between the slopes. The results showed only a significant trend for the difference in starting points between the negative and positive condition [$z - value = 1.697,\ p - value = 0.093$]. These results were in line with our hypotheses and were indeed very interesting, because they showed evidence for a clear distinction between these two conditions.

Regarding the slope differences, the results using the power law fit[11] are presented in table 4.3.

[11] Similar results were obtained using the exponential fit.

Conditions tested against each other		$z-value$	$p-value$
Approach Negative	Approach Neutral	2.622	**0.009**
Approach Negative	Avoid Positive	6.484	$<$ **0.001**
Approach Negative	Avoid Neutral	0.177	0.86
Avoid Neutral	Approach Neutral	2.45	**0.014**
Avoid Neutral	Avoid Positive	6.319	$<$ **0.001**
Avoid Positive	Approach Neutral	-3.854	$<$ **0.001**

Table 4.3: Post-hoc contrast performed on the difference between the slopes of different conditions, using the power law fit.

From the results of table 4.3 and figures 4.12 and 4.13, we inferred that the positive condition did not only start faster, but also seemed to have been easier to learn than the other conditions (steeper slope). On the other hand, the negative condition appeared to be the most difficult condition to be learned (less steepest slope). Besides that, it could also be inferred that the *approach negative* condition and the *avoid neutral* condition presented almost parallel trends, thus explaining the non-significant difference between the two slopes of these conditions.

4.2.4. Participants' ratings

After analyzing the results obtained from sub-section 4.2.3, we wanted to check whether the training had changed how participants perceived the pictures. Thereby, the hypotheses we had regarding changes in the valence ratings were quite straightforward:

- The first one was that participants would rate the positive pictures as positive, the negative ones as negative and the neutral ones, on average, as neutral, meaning they would be rated as zero. Additionally we also predicted that the rating of the pictures of the conditions that participants did not train would not change.
- During the period of training, in the negative group, we expected to observe an increase in the rating of negative pictures and a decrease in the rating of the neutral ones. In contrast, in the positive group, we expected an increase in the rating of neutral pictures and a decrease in the rating of positive pictures.
- Finally, similar to what was expected in sub-section 4.2.2, we predicted participants to generalize the effects of the training. Thereby, participants from the negative group would rate the negative pictures that they did not train higher than when they saw them for the first time. On the other hand, in the positive group we predicted the opposite, meaning that participants would rate the positive pictures they did not train lower than when they saw them for the first time.

In order to test the hypotheses mentioned above, two mixed models were designed: One regarding only the classifications of the trained pictures (equation 4.5):

$$classification = category * session + (1|subject). \quad (4.5)$$

And another one to analyze the differences between the first classification and the last one including the pictures used for generalization (equation 4.6):

$$classification = category * session * trained + generalized + (1|subject). \quad (4.6)$$

The analysis of the first mixed model shows that there is a highly significant main effect of category [$F_{(3,207.91)} = 306.78, p - value < 0.001$] and a trend for the main effect of session [$F_{(1,971.99)} = 3.42, p - value = 0.065$]. These results show that there are clear differences between the classifications of the different categories of pictures and that there might be a tendency towards change, along sessions. However, the interaction term did not present evidence of being significant [$F_{(3,971.99)} = 1.41,\ p - value = 0.240$], meaning we could not find evidence for a clear change over time for different categories.

Figure 4.14 depicts the evolution of the ratings, along sessions, of the pictures trained in the two groups.

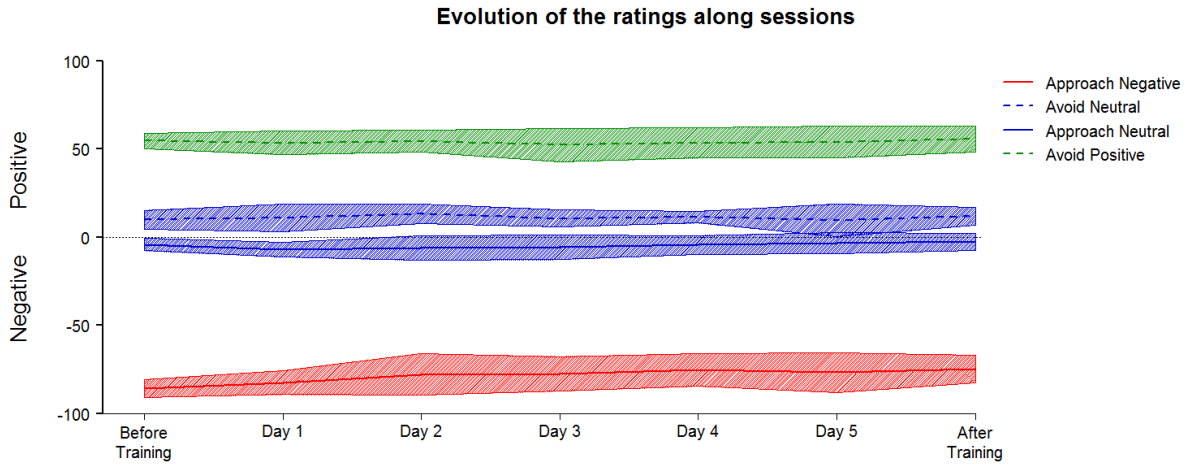
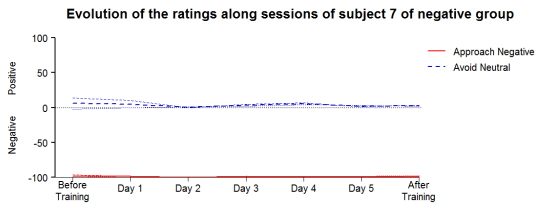
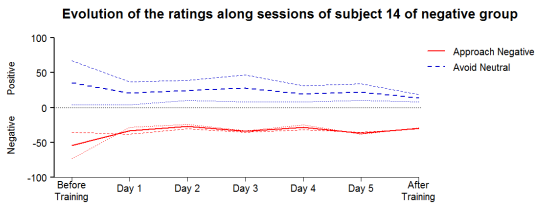


Figure 4.14: Evolution of the ratings along 5 days of training of the pictures trained in the 4 different conditions.

From figure 4.14, the difference between ratings of different categories is clear. Besides that, even though there is no significant variation of the green and the two blues lines, taking a closer look at figure 4.14 reveals a slight increase in the average of the ratings of the negative pictures. Considering the significant trend of the main effect of session, we decided to test for the significance of the slopes of the different conditions. The result confirms our exploratory visual analysis, *i.e.*, there was a significant increase regarding the rating of the negative pictures [$z - value = 2.650,\ p - value = 0.008$], while the others did not significantly change over time ($p - value > 0.200$). Moreover, we suspected that the analysis at the group level was possibly masking interesting results. In fact, looking at the individual level, we verified that there are two major types of results within each group. Figures 4.15a, 4.15b, 4.15c and 4.15d depict those results.





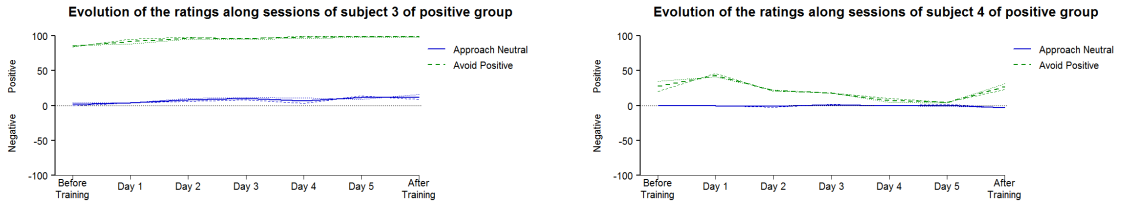


Figure 4.15: Evolution of the individual ratings along 5 days of training. (a) Depicts the evolution of the 7th participant of the negative group, (b) depicts the evolution of the 14th participant of the negative group, (c) depicts the evolution of the 3rd participant of the positive group and (d) depicts this evolution of the 4th participant of the positive group. The different less marked dashed lines correspond to the evolution of the trained pictures.

From figure 4.15, we conclude that some participants in the negative group did not change the rating of the negative pictures throughout the week, while others actually did. These two different outcomes might result from two different factors: the picture itself and the habituation to the picture.

Considering the reports that participants provided in the protocol file, there were participants that rated the picture according to the value inherent to it, even though they were developing feelings less aversive towards it. This might explain the cases of the 1st, 2nd, 4th, 5th, 7th, 8th, 10th and 16th participants of this group that reported constant ratings. On the other hand, the remaining participants of this group actually rated the pictures according to their habituation and changing feelings towards them, which in turn could explain the alterations in the ratings.

Regarding the positive group, it could be noticed that – besides of the effects reported for the negative group – there was another problem: there were participants who rated the positive pictures close to zero.

These results could be more accurate if we always asked participants to rate all the pictures. By doing this, we might disperse the effect of comparing neutral pictures against negative ones (for the negative group) and against positive ones (for the positive group). Besides, the question made regarding the way how participants should rate the pictures could be more in line with the one in [28], by explicating explicitly how they should rate the pictures.

Another interesting result was that, on average, participants rated the pictures of the condition *approach neutral* as negative. In line with these valence ratings, above, we show that participants have an avoidance bias towards the neutral pictures (cf. sub-section 4.2.2). This finding will be further discussed in the final chapter, but it supports our assumption that our task design was capable to depict that participants react according to their automatic tendencies.

Regarding the second mixed-model (cf. equation 4.6), whose variables' description was done in sub-section 4.2.2, its analysis showed a significant main effect of category [$F_{(3,1118.76)} = 168.78, p - value < 0.001$], a significant main effect of session [$F_{(1,1116)} = 5.05, p - value = 0.025$], a trend for the main effect of generalization [$F_{(1,1116)} = 2.93, p - value = 0.090$], but no main effect of trained vs. untrained pictures [$F_{(1,1116)} = 1.4, p - value = 0.240$]. These results were not only in line with the ones presented above, but also showed that there was a clear difference between the first and last time participants rated the pictures. Regarding the interaction terms, none of them presented any significance ($p - value > 0.200$).

Figure 4.16 depicts the ratings before and after the training of the positive group, towards negative and neutral pictures they did not train, and for the negative group, towards the positive and neutral pictures they did not train (left side), and ratings before and after the training of positive group, towards

the conditions they trained, and for the negative group, towards the conditions they trained, including the generalization results, respectively (right side).

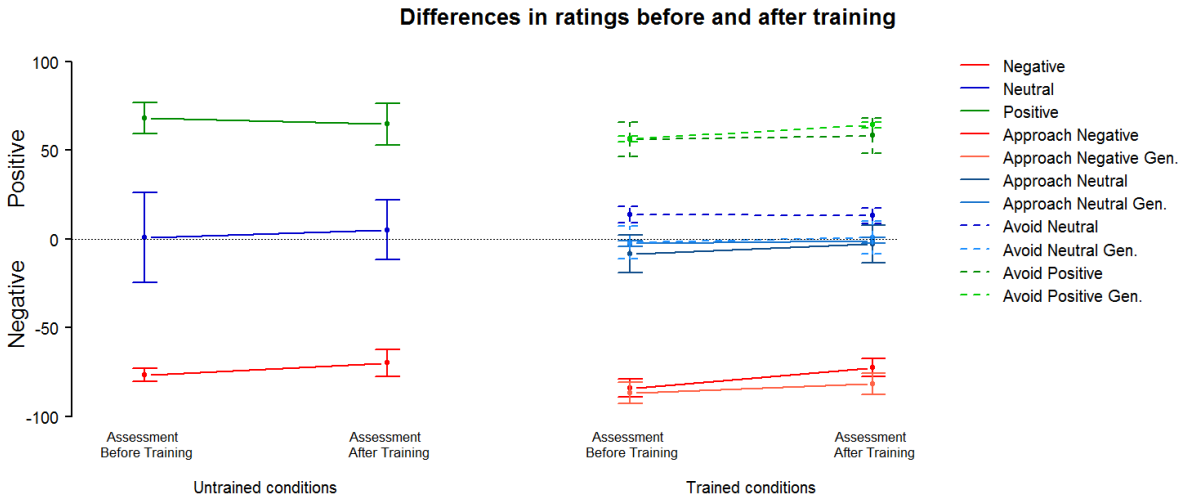


Figure 4.16: Ratings provided by the participants before and after the training of the positive group, towards negative and neutral pictures they did not train, and for the negative group, towards the positive and neutral pictures they did not train (left side), and ratings before and after the training of positive group, towards the conditions they trained, and for the negative group, towards the conditions they trained, including the generalization results, respectively (right side).

Considering the visual conclusions from figure 4.16 and the results provided by the analysis of equation 4.6, in order to verify whether the statistical results showed the same, post-hoc contrasts were computed. Considering the ratings **before the training** (left part in both sides of figure 4.16), we tested whether the positive ratings were significantly positive, whether the negative ratings were significantly negative and whether the neutral were not significantly different from zero. The results showed high significance (positive: $z - value = 13.87,\ p - value < 0.001$; negative: $z - value = 20.28,\ p - value < 0.001$; neutral: $z - value = -0.248,\ p - value = 0.800$). Moreover, there was no difference between the ratings **before the training** on left and right side of figure 4.16, confirming that participants perceived the pictures as expected.

Besides this, pair-wise comparisons were done between scores before and after the training, regarding the ratings of the trained and untrained conditions. The only significant result was the difference of the negative ratings (before vs. after training) for the trained conditions [$z - value = 2.11,\ p - value = 0.035$], which supported the abovementioned results and findings: Participants who trained to approach negative pictures showed alterations in their reactions to negative pictures over the time. All the other results did not reach significance ($p - value > 0.200$), as also did not the results for the generalization pictures of each category.

4.2.5. Practical test's behavioral data

To test whether the training was really effective, especially in what concerns changing the attitude towards negative and positive stimuli, a practical test was performed. In this test, participants of the two groups were separated into two further groups: one group had to sit on a pillow with a positive picture (cf. figure 3.6a sub-section 3.1.7) and the other one had to sit on a negative picture (cf. figure 3.6b sub-section 3.1.7). Following, we present our expectations for further discussions of the results:

- Within groups, we hypothesize that participants from the negative group will be faster to sit on the negative picture than on the positive picture. In the positive group, we expect the same pattern, although with a smaller difference.
- Between groups we expect participants form the negative group to sit down faster on the negative picture when compared to participants from the positive one sitting down on that same picture. The same pattern is expected regarding the positive picture.

Therefore, the mixed model approach was used and the following design was tested (equation 4.7):

$$rt = group * pillow\_picture + (1|subject). \quad (4.7)$$

The analysis of the designed model revealed that there was no evidence of any significant main effects, nor their interaction (all $p - value > 0.200$).

Figure 4.17 depicts the mean RTs in seconds (s) that participants took to sit down on a specific kind of picture for each group.

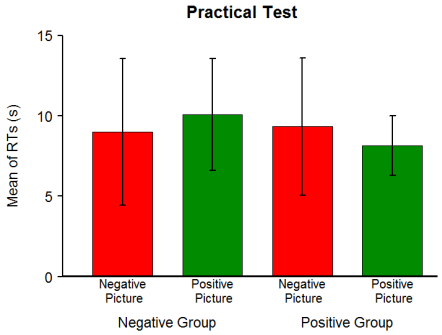


Figure 4.17: Results from the practical test, which have depicted the mean RTs in seconds (s) that participants took to sit down on a specific category of picture for each group.

Although there is no significant main effect, figure 4.17 depicts some of our hypotheses such as the participants from the negative group being on average faster when sitting down on the negative picture than when sitting down on the positive picture. In fact, they are, even though only slightly, faster when compared to participants from the positive group who had to sit down on the negative picture. Still, not much weight can be given to these results at the moment.

Deeper discussions, specifically of the shortcomings of this practical test will be given in the final chapter.

4.3. Model-based analysis

In this section, we will describe the model based analysis of the behavioral data concerning the training version of the AAT of the second sample.

The data used for fitting the computational models was treated according to what was described in sub-section 4.2.3. The procedure of joining the 5 sessions' data as described in sub-section 4.2.3 was also helpful for this analysis, due to the considerable high number of estimated parameters and due to the multicollinearity[12] problems we anticipated from the analysis of equations 3.7 and 3.8 (cf.

[12] High correlation between the explanatory variables.

respectively section 3.5 and sub-section 3.5.1): According to Shen (1998), we should have at least 20 to 30 observations per parameter estimated [121].

The first objective of this analysis was to verify if there were significant differences in model frequencies within the population, and if there was one type of models that explained the data better than the remaining ones. Moreover we were interested in finding if our computational models were able to explain the subject's behavior significantly better than models which did not assume a dynamic change of the RTs with the trial number, as a function of the interaction between the habit learning, Pavlovian and cognitive control components.

The comparison between the models described in table 3.2 (cf. sub-section 3.5.3) was performed through random effects Bayesian model selection (BMS). To proceed with this, first we used a slightly modified version of the toolbox indicated in [116] to perform all Bayesian model comparisons (BMCs). This allowed us to compute, besides the Exceedance Probabilities (EPs), also the Protected Exceedance Probabilities (PEPs). In order to execute the BMC, we provided the above mentioned toolbox with an approximation of the model evidence (ME) of the different models, the BIC's. These values were computed according to equation 3.23, through the maximum-likelihood estimation method (cf. section 3.6) using the variances estimated according to equation 3.16 for the Normal likelihood function [$RT_i \sim Normal(\widehat{RT_i}, \sigma^2)$] and to equation 3.20 for the Log-Normal likelihood function [$RT_i \sim LogNormal(Log(\widehat{RT_i}), \sigma^2)$].

Consequently the BMS was executed through the analysis of the PEPs obtained for the models specified in table 3.2, whose significance threshold was defined at 95%.

Initially, we started to analyze directly the RTs and then applied the moving average filter (as described in sub-section 3.5.3) and proceeded in a similar way.

4.3.1. Analysis of the RTs

Following the hypothesis mentioned in sub-section 3.4.2, we first aimed to verify whether the predicted RTs would or would not be normally distributed. Therefore we performed the maximum likelihood estimation through the Normal and the Log-Normal likelihood functions for all the designed models, as described in table 3.2.

With exception of the first two models, all the remaining had the same number of parameters, therefore the use of the BIC as an approximation of the model evidence was considered reasonable. The outcome of the BMC showed highly significant results since a PEP over 95% was obtained for the fourth model.

Figure 4.18 depicts the outcome of the BMC between the models described in table 3.2.

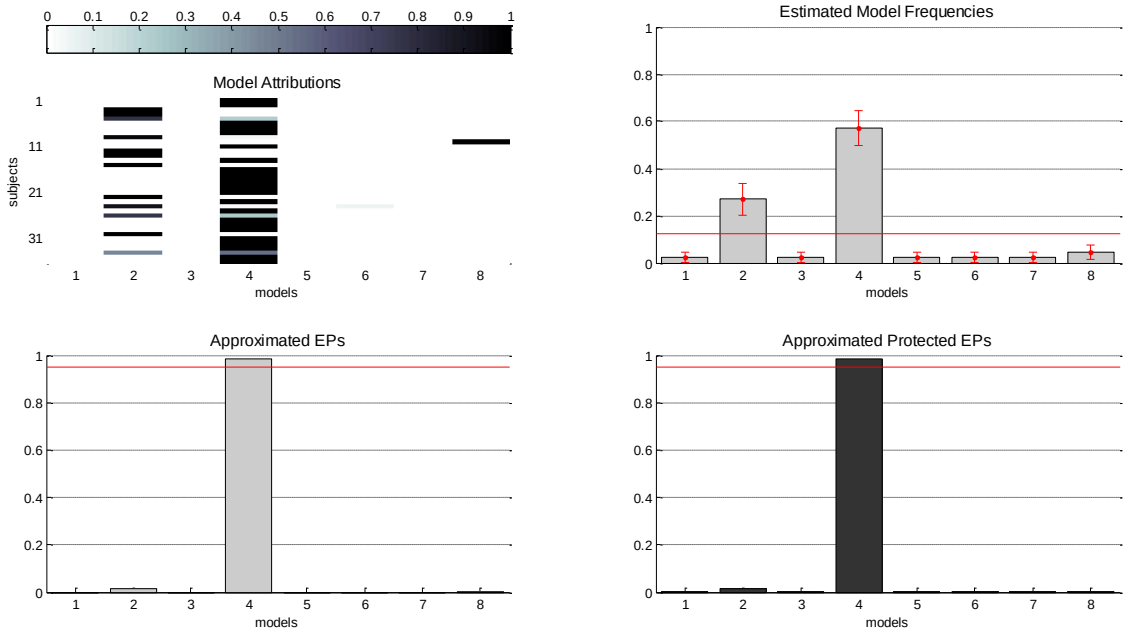


Figure 4.18: Results of BMC between the models described in table 3.2 obtained using the BIC, for the population. (Top Left) Model Attribution at subject level, the first eighteen subjects belong to negative group while the last eighteen to the positive group. (Top Right) Estimated model frequencies. (Left and Right Bottom) Approximated Exceedance Probabilities (EPs) and Protected EPs of the eight models.

From figure 4.18 we could corroborate the hypothesis that the observed RTs were not normally distributed, even when considering short portions of the training period. These results were also in line with the hypothesis that the initially designed model was able to describe the subjects' behavior and (partially) captured the psychological and cognitive processes we wanted. Besides this, these results also showed evidence that the Hebbian and non-Hebbian components (associated to learning rates α and β respectively) were necessary to describe subject's habit learning.

Then we proceeded with the analysis of the significance of the Pavlovian parameter. This analysis was performed considering the parameters estimated for all subjects through the model revealing the higher value of PEPs.

In order to perform a significance test, we first had to check whether this parameter's sample was or was not normally distributed, because the test to be used to verify the null hypothesis (that the parameter was significantly different than zero) depended on that prerequisite [122]. Therefore, we resorted to *Kolmogorov-Smirnov* test since it was proven to be one of the most powerful normality tests [123]. The result showed that we should reject the null hypothesis (the π being normally distributed) with high significance ($p - value < 0.001$). Consequently, we performed a Wilcoxon signed rank test which does not make any distributional assumptions to test if the Pavlovian parameter was significantly different than zero [112]. Even though the result was not significant, it presented a trend ($p - value = 0.066$).

Although, considering figure 4.18, we guessed that this analysis included participants that should not be considered, *i.e.*, there was the possibility of having outliers' participants.

Therefore we decided to perform a histogram of the parameter's sample (figure 4.19).

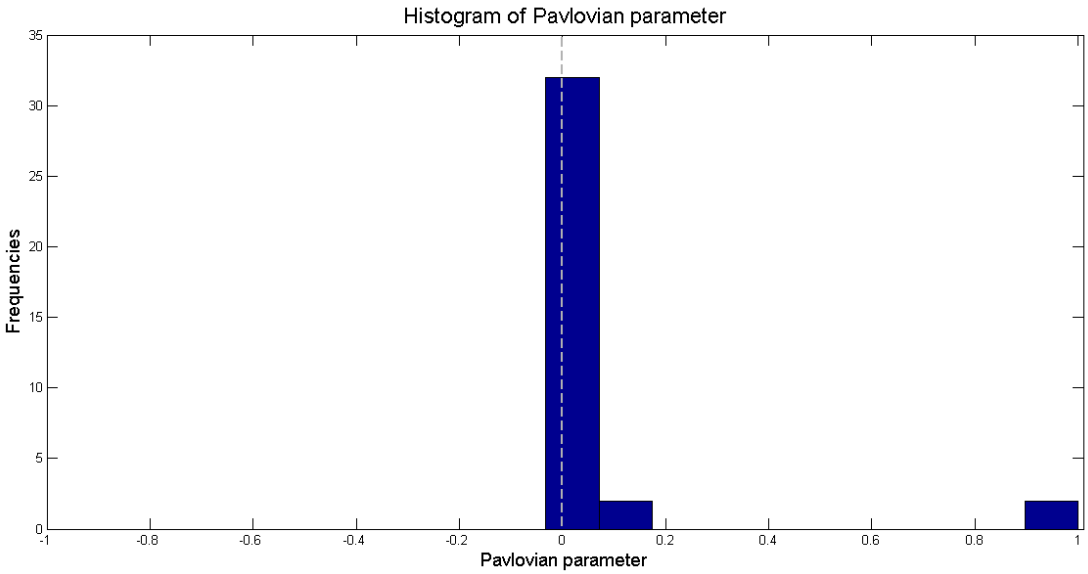


Figure 4.19: Histogram of the Pavlovian parameter estimated by the computational model selected through BMS. The vertical dashed grey line identifies the value zero.

Figure 4.19 confirmed that we had (at least) two outlier candidate subjects. Thus, we proceeded to verify which subjects had been assigned with a Pavlovian parameter close to 1. We concluded that that had been the case for the 4th and the 8th subjects of the positive group (subjects 22 and 26 respectively).

Due to the constraint imposed on the cognitive control component (cf. equation 3.8 in sub-section 3.5.1), we expected a high multicollinearity between these two components, which, from the analysis of the equation that rules the preferences (cf. equation 3.2 section 3.5), could only be dissipated if participants had rated at least three of the four pictures trained or two of the four (if these two were of different conditions) differently from zero. This was, because, if only two pictures (of the same category) were rated differently than zero, both the Pavlovian and cognitive control component would be estimated only from the data of that condition, leading to a huge increase of the covariance between them: Since they were only influenced by one condition, as long as the difference between these two components was kept approximately constant, all estimated values for the parameters would work "equally" well. Therefore we verified the individual ratings of participants referred and observed that subject 22 was in the condition above mentioned. Additionally, we verified that this subject's behavior was better described by the simplest model (where the mean and not the decision making system-related components were modeled).

Figure 4.20 depicts the predicted RTs for 4th subject of the positive group (subject 22).

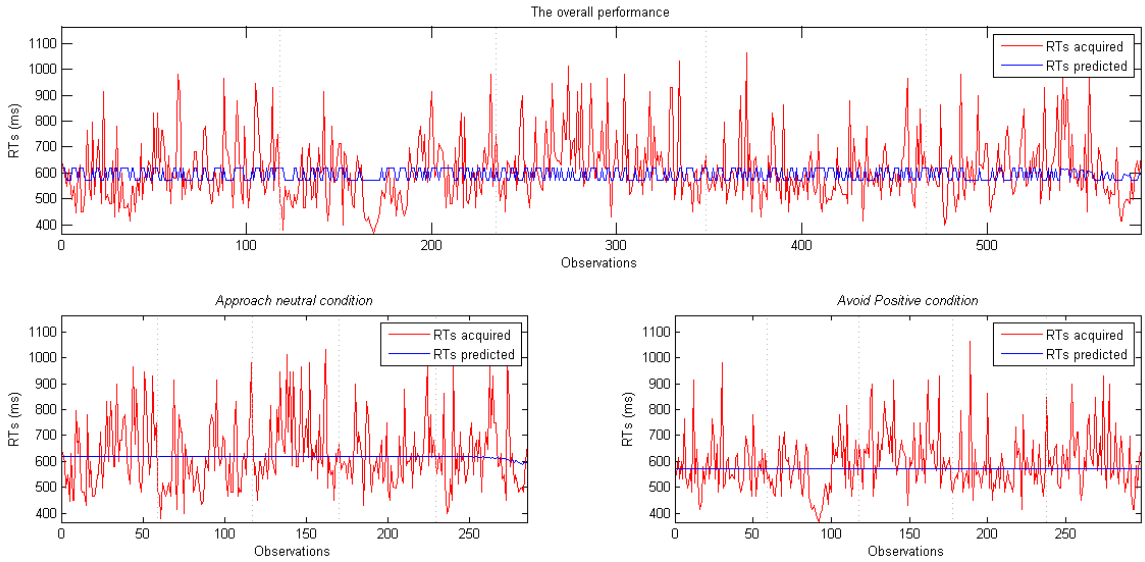


Figure 4.20: RTs predicted by the computational model for 4th subject of the positive group. The red line represents the RTs acquired and the blue line the RTs predicted by the computational model.

Figure 4.20 confirmed the result observed in figure 4.18.

Following this reasoning, we verified the ratings of every subject and found that the 5th, the 6th, and the 9th subjects of the positive group (subjects 23, 24 and 27 respectively) had the same problem. Therefore, we tested again if the Pavlovian parameter was significantly different from 0, excluding the parameters of subjects 22, 23, 24 and 27. The result showed that the Pavlovian parameter was not significant ($p - value = 0.16$). Although that was expected since the result from subject 22 was coincidently improving the previous test.

Following our *a priori* hypotheses, we then proceeded to the comparison of the computational models which received as input transformed RTs (via moving average filtering).

Motivated by the intra-subject variability verified in the results of the arrow version of the AAT (cf. appendix A4) and by the procedure performed by Palminteri *et al.* (2011), in order to reduce the noise caused by undesired processes (cf. section 3.3), we applied the moving average method to each subject's data according to what as described in sub-section 3.5.1. Then we followed a procedure similar to the one explained above.

Although, the results of the BMC did not show that any model had reached the significance threshold, since the originated PEPs were not over 95%, we could actually observe that the tendency of the results presented above did not change (cf. figure 4.21).

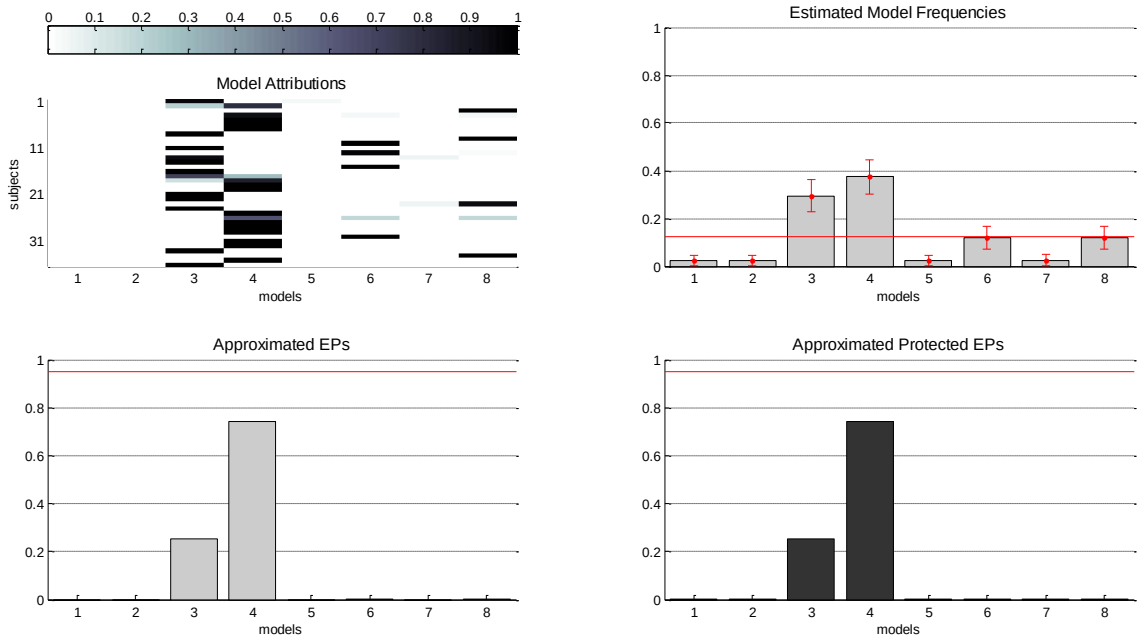


Figure 4.21: Results of BMC between the models which received as input transformed RTs (via moving average filtering) described in table 3.2 obtained using the BIC, for the population. (Top Left) Model Attribution at subject level, the first eighteen subjects belong to negative group while the last eighteen to the positive group. (Top Right) Estimated model frequencies. (Left and Right Bottom) Approximated Exceedance Probabilities (EPs) and Protected EPs of the eight models.

From figure 4.21 we observed that the number of subjects whose parameters were estimated by our model using the normal likelihood function increased a lot relatively to figure 4.18. This resulted from the application of the moving average method since it reduced not only the noise, but also attenuated the effects of longer RTs. Nonetheless, comparing figures 4.18 and 4.21, we noticed that the increase of subjects assigned to model 3 resulted mainly from subjects that were before assigned to one of the simplest models.

The increase of subjects assigned to model 6 and 8 was also a consequence of the application of the moving average method. Although, this increase was not significant since it did not reflect any change in the originated PEPs from this models.

Therefore, we considered the analysis of the BMC including only the first four models (cf. appendix A5) and this procedure resulted in the increase of the PEPs to 90% for model 4.

Then, we proceeded with the statistical analysis of the Pavlovian parameter, following the procedure explained above. First we tested the Pavlovian parameters' sample for normality using the *Kolmogorov-Smirnov* test and the result was in line with the one obtained above ($p - value < 0.001$). Consequently we used the Wilcoxon signed rank test to assess the significance of this parameter. The result, although better, only showed again a trend ($p - value = 0.061$). However this test was under the influence of subjects 22, 23, 24 and 27. After excluding them from the analysis, the result became significant ($p - value = 0.036$). In addition, we performed a histogram similar to the one depicted in figure 4.19 (figure 4.22).

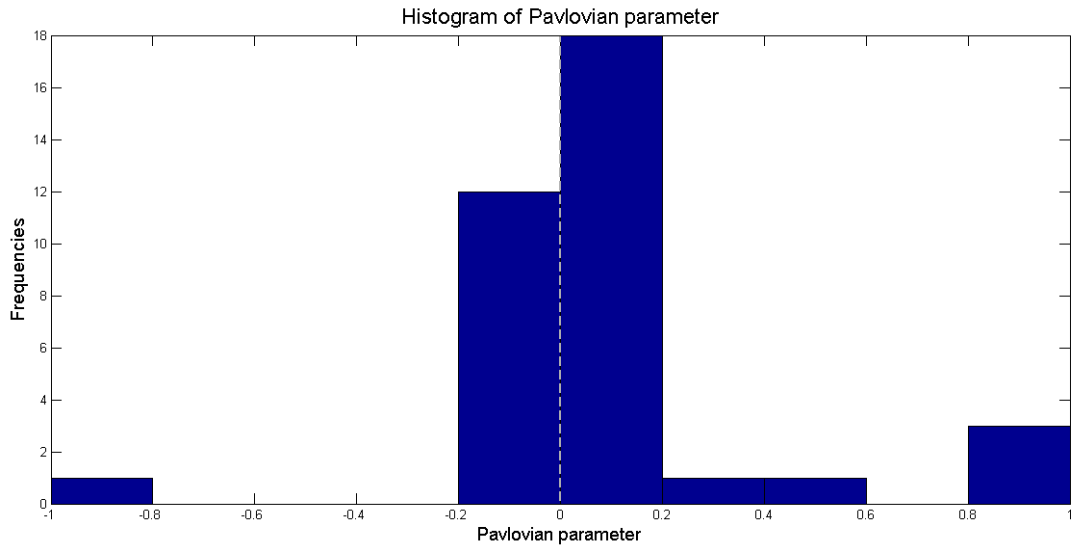


Figure 4.22: Histogram of the Pavlovian parameter estimated by the computational model selected through BMS, considering the application of the moving average method on the raw data. The vertical dashed grey line identifies the value zero.

From figure 4.22 we corroborated the finding that the subject 22's estimate of the Pavlovian parameter of the positive group was not reliable (since this was the subject who had the *-1* value of the Pavlovian parameter assigned to).

Therefore, the significance achieved on the Pavlovian parameter can be explained not only by the exclusion of the four referred subjects, but also by the increase of the Pavlovian parameter's value for other subjects.

This result was in line with the hypothesis that by applying the Moving Average we would reduce the noise of the undesired processes referred in section 3.3.

5

Conclusions and Future Developments

Contents

5.1. Conclusions

This thesis aimed at understanding the basic mechanisms that underlie the overcoming of automatic response tendencies on healthy subjects through standard analysis of behavioral data and the development of novel neuro-computational models. In order to achieve this, we developed and applied different versions of the Approach Avoidance Task. These were performed by 36 healthy subjects, who were divided in two distinct groups (the negative group and the positive group).

The analysis of the behavioral data was divided according to the different versions of the task and the subsequent conclusions will follow the same structure.

From the arrow version of the AAT we conclude that, although there are intra-individual differences regarding the two actions, at a group level these differences are not significant. It is also important to refer that we questioned participants if they thought there was any movement easier to perform compared to the other. Even though no rigorous analysis was performed, we confirmed that half of the participants that reported one movement to be easier than the other were faster performing the opposite movement.

The assessment version was created in order to measure the unintentional valence processing of each subject, before any exposure to the stimuli they would train and secondly, after the train they went through during 5 consecutive days. From the analysis of sub-section 4.2.2, we verified that participants from the negative group presented an increase in their approach bias towards negative pictures, while participants from the positive group presented a slightly increase in their avoidance bias towards positive pictures. Besides this, we also verified that participants from the negative group showed evidence for generalization effects regarding the negative pictures. Therefore, we concluded that these findings might be a consequence of habit formation imposed to participants that was interpreted by the significant decrease of the RTs.

Nevertheless, these results cannot be considered to be outstanding, due to factors such as the great inter-subjects variability and some problems inherent to this version of the task. In what concerns the first factor, it is important to emphasize that RTs are a very difficult behavioral measure to deal with, even after the pre-processing we performed (cf. section 3.3). Even though we put considerable effort into the development of this version and actually the majority of participants reported that they simply focused on the arrows and not on the images along presented, we found problems with this version. Therefore, we reach the conclusion that it must be modified, because the simplicity of the design allowed that undesired processes interfered with the effect we desired to measure. Notwithstanding, Wiers *et al.* (2009) showed that participants, who were instructed to react to the format of pictures, were influenced by the content of the pictures [83]. Following this reasoning, although our design has reduced the WM load by a considerable amount comparing with Wiers' study, it seems we could still (partially) capture some of the desired effects.

Another finding supporting that the assessment version still captured the effect mentioned was the fact that participants had an initial bias towards neutral pictures that was in line with their ratings. More specifically, the neutral pictures used for the *approach condition* (buses and city related pictures) were rated negatively and, as we proved, participants had an avoidance bias towards them.

Nonetheless, this version was also designed considering a different audience: children. Consequently, the modifications that might be done to it require caution to not change the design in such a way that it becomes too difficult for this future audience.

Then, direct analysis of the behavioral data of the training version of the AAT showed that participants of both groups learned the trained conditions. This finding shows evidence that repetition actually plays a key role in habits formation [69] [71] [72]. A more detailed analysis also showed that there is a clear difference between the two incongruent conditions which were tested (approach negative and avoid positive). This difference in difficulty of learning the *approach negative* condition over *avoid positive* condition might result from the fact that negative pictures were considered more negative than the positive pictures positive. This statement was statistically evident since we performed a post-hoc contrast between the absolute ratings of negative pictures and positive picture before the training ($z - value = 5.353, p - value < 0.001$). Therefore, according to Cacioppo *et al.* (1997), the negative pictures probably elicited stronger avoidance tendencies than the positive pictures did elicit approach tendencies. Consequently inhibiting such avoidance tendencies in incompatible conditions was much more cognitively effortful than inhibiting positive approach tendencies [120].

Another important finding, which might be a consequence of habit formation, as transduced by the significant decrease of reaction times of participants of the negative group throughout the training period, was the built-up association between the negative stimuli and the approach reactions. This assumption was supported by the significant difference verified between the first and the last time participants rated the pictures they had trained (cf. sub-section 4.2.4).

Regarding the practical test we did not explore more the results because we lacked a group of control to compare the results obtained with. Therefore, we cannot validate the results, even though visual inspection revealed that participants from the negative group seemed to sit (on average) faster on the negative picture than on the positive one, in line with our hypothesis.

Nonetheless, we hypothesize that in the control group participants will sit faster on the positive picture than on the negative picture. This hypothesis is based on fact that participants usually tend to approach positive stimuli and avoid negative ones [96]. In addition, we expect people who have trained to approach negative stimuli to be faster at sitting on the pillow with the negative picture and to have a similar behavior regarding the positive picture. On the other hand, we expect people who have trained to avoid positive stimuli to be slower at sitting on the pillow with the positive picture and to present a similar behavior regarding the negative picture.

To study the psychological and cognitive processes of the training behavioral data, a novel computational learning framework was also developed. Furthermore, due to the uncertainty regarding the habit learning strategies which could have been followed by the different subjects, we modeled three alternatives ways (cf. table 3.2 in sub-section 3.5.3).

This approach allowed us to test whether participants, in general, presented or not evidence for Hebbian learning, which was associated to the multiplicative learning rate α, or alternatively, if both Hebbian and non-Hebbian components were necessary to describe subject's habit learning.

Therefore, a Bayesian model comparison (BMC) framework was used for model selection. This was a very useful tool, since it allowed all hypotheses to be tested, without requiring prior assumptions on

specific distributions of the models or the parameters on the populations to be made. To perform this process we resorted to model evidence's approximation, the Bayesian Information Criterion. The BIC, despite some inaccuracies is a reasonable (and widely-used) measure.

In the overall model comparison process, we observed that the initial model we implemented yielded the best results for the majority of the participants. This indicated that participants, in general, needed the two components of learning. In fact this result was in line with the hypothesis that without the Hebbian component there would not be a strengthening of the synaptic efficacy that arose from the presynaptic cell's repeated and persistent stimulation of the postsynaptic cell [99], and without the other component of learning we would not be able to capture the learning in the early stage of the task, where a purely Hebbian framework would be much more ineffective since there were no previous cumulative experience of the *stimulus-response* pairing.

Moreover, the BMC also provided findings indicating that the RTs were not normally distributed. Even the application the MA method did not change this trend.

In fact the application of this method proved itself to be good idea, because, even though the moving average method might have attenuated some processes of interest, we were able to reduce the noise from undesired cognitive and motor processes. This fact became relevant when we observed the increase in significance of the Pavlovian parameter.

More important, though, was the fact that we found consistent and coherent findings in both of the performed analyses. Therefore, this thesis provides evidence that internal conflicting mechanisms exist when performing incongruent conditions and that the participants subjected to the 5 consecutive-day protocol were able to learn the conditions trained through habit formation mechanisms.

5.2. Future work

As it was demonstrated, very interesting findings supported several hypotheses we had. Nonetheless, throughout the performed analyzes we also referred to possible problems that are important to take into account in further studies.

First, it is important to refer that it was deceiving to not found a good database with validated pictures that could fulfill our requirements. Even though we did not have a real problem in finding it, we think that before using different pictures on this task it is important to validate the pictures used. Otherwise we might occur in the mistake of using pictures that are very subjective and this could have negative repercussion on the final results, especially on the model-based results that use this information directly.

Regarding the design of the assessment version we also think that it should be modified in order to increase the WM load, since apparently this design reduced it considerably. Notwithstanding the use of a dual processing task[13] might also help because it would not only reduce the noise of the other cognitive processes through attentional allocation, but also would allow to better measure the automatic reactions and consequently the processes that elicit them.

Besides that we should had instructed the participants without presenting explicitly the stimuli, since we might have allowed participants to prepare for the conditions they had to train, especially the

[13] Tasks in which the participant, besides performing the normal AAT, would also perform another task that would require the participant to be more focused when performing these routines.

incompatible ones. Therefore this fact might have enhanced the cognitive effort when performing the task thus contributing to the dissimulation of the influences from automatic processes [19].

Another task-related problem that we think it should have been more clear was the way participants were rating the pictures, *i.e.*, the question we made should be more clear in terms of what people were really feeling about the picture and not just in terms of the content of the picture. Besides, in order to have a more reliable evaluation, we should ask participants to rate all the pictures at the same time, since if that is the case we can really assess the effects on the trained pictures compared to the ones used for generalization.

Regarding the computational model, for the scope of this thesis, we assumed the Pavlovian and cognitive control components as constant parameters. This allowed us to capture psychological and cognitive processes which allowed us to understand and provided us insight regarding the importance of them. Nonetheless, much more effort is required for the model to better explain the behavioral data. This might be accomplished by taking into account more complex formulas describing the dynamics of the Pavlovian and cognitive control components.

Still, they were not considered because our model is already able to transduce the prevalence of habit learning component over the two other components, which values would decrease because the neuronal activation is no longer required [67].

Concerning the multicollinearity problems verified in sub-section 4.3.1, we thought that this problem could be slightly reduced if the cognitive control parameter was always present or even if we use another component related to the cognitive control. In fact, this idea is actually worth trying, since there has been evidence that during the performance of similar versions of the AAT it was been hypothesized that the WM component is always present.

Following this reasoning, the implementation of hierarchical models of parameter estimation could be much more advantageous, since the hierarchical models cope with within-subject variability on parameter estimates, and this variability might be problematic in situations where there are subjects whose data is not well-explained by the selected model.

Regarding model comparison, the use of better model evidence's approximations (computed through Markov Chain Monte Carlo sampling [105] or variational Bayes techniques [113] [115]) should also be targeted, in order to formally validate model selection procedures. Besides, although this approach requires a long execution time and it is extremely costly from a computational perspective, we already started to develop this methodology due to the fact that it is much more reliable than the methods of optimization.

Although it is not possible to perform invasive electrophysiological recordings in humans, like it was done in animals, indirect measures could also be done, like the Blood-Oxygen Level Dependent (BOLD). So, performing the task inside an fMRI scanner and performing an fMRI model-based analysis could also provide more information about the neuronal assumptions we made.

Moreover it might also allow to improve the model-based approach by providing insights regarding the neuronal structures involved when performing the AAT, since it actually is prepared for this purpose. Indeed, that is the plan for a near future, so that we can then apply this routines to OCD patients and develop an add-on therapy based on these routines.

R

References

[1] Williams, M. T., Farris, S. G., Turkheimer, E. N., Franklin, M. E., Simpson, H. B., Liebowitz, M., & Foa, E. B. (2014). The impact of symptom dimensions on outcome for exposure and ritual prevention therapy in obsessive-compulsive disorder. *Journal of anxiety disorders*, *28*(6), 553-558.

[2] Torp, N. C., Dahl, K., Skarphedinsson, G., Thomsen, P. H., Valderhaug, R., Weidle, B., ... & Ivarsson, T. (2015). Effectiveness of cognitive behavior treatment for pediatric Obsessive-Compulsive Disorder: Acute outcomes from the Nordic long-term OCD treatment study (NordLOTS). *Behaviour research and therapy*, *64*, 15-23.

[3] Piacentini J, Langley A, & Roblek T (1997). *Cognitive behavioral treatment of childhood OCD - It's only a false alarm*. New York, NY: Oxford University Press.

[4] Gillan, C. M., Papmeyer, M., Morein-Zamir, S., Sahakian, B. J., Fineberg, N. A., Robbins, T. W., & de Wit, S. (2011). Disruption in the balance between goal-directed behavior and habit learning in obsessive-compulsive disorder. *American Journal of Psychiatry*, 168(7), 718-726.

[5] Gillan, C. M., Apergis-Schoute, A. M., Morein-Zamir, S., Urcelay, G. P., Sule, A., Fineberg, N. A., ... & Robbins, T. W. (2015). Functional neuroimaging of avoidance habits in obsessive-compulsive disorder. *American Journal of Psychiatry, 172*(3), 284-293.

[6] Gillan, C. M., Morein-Zamir, S., Urcelay, G. P., Sule, A., Voon, V., Apergis-Schoute, A. M., ... & Robbins, T. W. (2014). Enhanced avoidance habits in obsessive-compulsive disorder. *Biological psychiatry, 75*(8), 631-638.

[7] Pauls, D. L., Abramovitch, A., Rauch, S. L., & Geller, D. A. (2014). Obsessive-compulsive disorder: an integrative genetic and neurobiological perspective. *Nature Reviews Neuroscience, 15*(6), 410-424.

[8] Maia, T. V., Cooney, R. E., & Peterson, B. S. (2008). The neural bases of obsessive–compulsive disorder in children and adults. *Development and psychopathology, 20*(04), 1251-1283.

[9] Carver, C. S. (2006). Approach, avoidance, and the self-regulation of affect and action. *Motivation and Emotion, 30*, 105-110.

[10] Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review, 8*(3), 220-247.

[11] Hofmann, W., Friese, M., & Strack, F. (2009). Impulse and self-control from a dual-systems perspective. *Perspectives on Psychological Science, 4*(2), 162-173.

[12] Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature neuroscience*, *14*(2), 154-162.

[13] Doya, K. (2007). Reinforcement learning: Computational theory and biological mechanisms. *Future*, *31*.

[14] Ludvig, E. A., Bellemare, M. G., & Pearson, K. G. (2011). A primer on reinforcement learning in the brain: Psychological, computational, and neural perspectives. *Computational neuroscience for advancing artificial intelligence: Models, methods and applications*, 111-144.

[15] Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of neuroscience*, *16*(5), 1936-1947.

[16] Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of neurophysiology*, *80*(1), 1-27.

[17] Friese, M., Hofmann, W., & Wanke, M. (2008). When impulses take over: moderated predictive validity of explicit and implicit attitude measures in predicting food choice and consumption behaviour. *British Journal of Social Psychology, 47*(3), 397-419.

[18] Hofmann, W., Gschwendner, T., Friese, M., Wiers, R. W., & Schmitt, M. (2008). Working memory capacity and self-regulatory behavior: toward an individual differences perspective on behavior determination by automatic versus controlled processes. *Journal of Personality and Social Psychology, 95*(4), 962-977.

[19] Krieglmeyer, R., & Deutsch, R. (2010). Comparing measures of approach–avoidance behaviour: The manikin task vs. two versions of the joystick task.*Cognition and Emotion*, *24*(5), 810-828.

[20] Solarz, A. K. (1960). Latency of instrumental responses as a function of compatibility with the meaning of eliciting verbal signs. Journal of experimental psychology, 59(4), 239. [21] Dickinson, Anthony. "Actions and habits: the development of behavioural autonomy." *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 308.1135 (1985): 67-78.

[21] Ernst, L. (2013). *Approaching the negative is not avoiding the positive: FNIRS, ERP and fMRI studies on the approach-avoidance task* (Doctoral dissertation, Universität Tübingen).

[22] Ernst, L. H., Plichta, M. M., Dresler, T., Zesewitz, A. K., Tupak, S. V., Haeussinger, F. B., Fischer, M., Polak, T., Fallgatter, A. J. & Ehlis, A.-C. (2014). Prefrontal correlates of approach preferences for alcohol stimuli in alcohol dependence. *Addiction Biology, 19(3),* 497-508.

[23] Wiers, R. W., Eberl, C., Rinck, M., Becker, E. S., & Lindenmeyer, J. (2011). Retraining automatic action tendencies changes alcoholic patients' approach bias for alcohol and improves treatment outcome. *Psychological Science, 22*(4), 490-497.

[24] Zhou, Y., Li, X., Zhang, M., Zhang, F., Zhu, C., & Shen, M. (2011). Behavioural approach tendencies to heroin-related stimuli in abstinent heroin abusers. *Psychopharmacology (Berl), 221*(1), 171-176.

[25] Heuer, K., Rinck, M., & Becker, E. S. (2007). Avoidance of emotional facial expressions in social anxiety: The Approach-Avoidance Task. *Behav Res Ther, 45*(12), 2990-3001.

[26] Eberl, C., Wiers, R. W., Pawelczack, S., Rinck, M., Becker, E. S., & Lindenmeyer, J. (2013). Approach bias modification in alcohol dependence: do clinical effects replicate and for whom does it work best? *Developmental Cognitive Neuroscience, 4*, 38-51.

[27] Najmi, S., Kuckertz, J. M., & Amir, N. (2010). Automatic avoidance tendencies in individuals with contamination-related obsessive-compulsive symptoms. *Behaviour research and therapy*, 48(10), 1058-1062.

[28] Amir, N., Kuckertz, J. M., & Najmi, S. (2013). The effect of modifying automatic action tendencies on overt avoidance behaviors. *Emotion*, 13(3), 478.

[29] Sharbanee, J. M., Hu, L., Stritzke, W. G., Wiers, R. W., Rinck, M., & MacLeod, C. (2014). The effect of approach/avoidance training on alcohol consumption is mediated by change in alcohol action tendency. *PloS one*, *9*(1).

[30] Radke, S., Güths, F., André, J. A., Müller, B. W., & de Bruijn, E. R. (2014). In action or inaction? Social approach–avoidance tendencies in major depression.*Psychiatry research*, *219*(3), 513-517.

[31] Ernst, L. H., Plichta, M. M., Lutz, E., Zesewitz, A. K., Tupak, S. V., Dresler, T., Ehlis, A.-C. & Fallgatter, A. J. (2013). Prefrontal activation patterns of automatic and regulated approach-avoidance reactions - A functional near-infrared spectroscopy (fNIRS) study. *Cortex, 49(1),* 131-141.

[32] Maia, T. V. (2009). Reinforcement learning, conditioning, and the brain: Successes and challenges. *Cognitive, Affective, & Behavioral Neuroscience, 9*(4), 343-364.

[33] Doya, K. (2007). Reinforcement learning: Computational theory and biological mechanisms. *future*, *31*.

[34] Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1, No.1). Cambridge: MIT press.

[35] Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275:1593-1599.

[36] Reynolds, J. N. J. and Wickens, J. R. (2002). Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks*, 15:507-521.

[37] Palminteri, S., Lebreton, M., Worbe, Y., Hartmann, A., Lehéricy, S., Vidailhet, M., ... & Pessiglione, M. (2011). Dopamine-dependent reinforcement of motor skill learning: evidence from Gilles de la Tourette syndrome. *Brain*, *134*(8), 2287-2301.

[38] Neumann, R., & Strack, F. (2000). Approach and avoidance: the influence of proprioceptive and exteroceptive cues on encoding of affective information. *Journal of personality and social psychology*, *79*(1), 39.

[39] Hall, J. E. (2010). Guyton and Hall textbook of medical physiology. Elsevier Health Sciences.

[40] Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1990). Emotion, attention, and the startle reflex. *Psychological review*, *97*(3), 377.

[41] Lang, P. J. (1995). The emotion probe: studies of motivation and attention. *American psychologist*, *50*(5), 372.

[42] Cacioppo, J. T., Priester, J. R., & Berntson, G. G. (1993). Rudimentary determinants of attitudes: II. Arm flexion and extension have differential effects on attitudes. *Journal of personality and social psychology*, *65*(1), 5.

[43] Markman, A. B., & Brendl, C. M. (2005). Constraining theories of embodied cognition. *Psychological Science*, *16*(1), 6-10.

[44] Lang, P. J., & Davis, M. (2006). Emotion, motivation, and the brain: reflex foundations in animal and human research. *Progress in brain research*, *156*, 3-29.

[45] Ernst, M., & Fudge, J. L. (2009). A developmental neurobiological model of motivated behavior: anatomy, connectivity and ontogeny of the triadic nodes.*Neuroscience & Biobehavioral Reviews*, *33*(3), 367-382.

[46] Damasio, H., Grabowski, T., Frank, R., Galaburda, A. M., & Damasio, A. R. (1994). The return of Phineas Gage: clues about the brain from the skull of a famous patient. Science, 264(5162), 1102-1105.

[47] Seger, C. A., & Spiering, B. J. (2011). A critical review of habit learning and the Basal Ganglia. *Frontiers in systems neuroscience*, *5*.

[48] Du, Z. (2014). *Caractérisation of GABAergic neurotransmission within basal ganglia circuit in R6/1 Huntington's disease mouse model* (Doctoral dissertation, Bordeaux).

[49] Conceição, V. M. A. (2014). *Study of habit learning impairments in Tourette syndrome and obsessive-compulsive disorder using reinforcement learning models*. (Master's Thesis, Instituto Superior Técnico, Lisbon, Portugal).

[50] Phillips, M. L., Ladouceur, C. D., & Drevets, W. C. (2008). A neural model of voluntary and automatic emotion regulation: implications for understanding the pathophysiology and neurodevelopment of bipolar disorder. *Molecular psychiatry*, *13*(9), 833-857.

[51] Davidson, R. J., Jackson, D. C., & Kalin, N. H. (2000). Emotion, plasticity, context, and regulation: perspectives from affective neuroscience. *Psychological bulletin*, *126*(6), 890.

[52] Wager, T. D., Phan, K. L., Liberzon, I., & Taylor, S. F. (2003). Valence, gender, and lateralization of functional brain anatomy in emotion: a meta-analysis of findings from neuroimaging. *Neuroimage*, 19(3), 513-531.

[53] Lang, P. J., Cuthbert, B. N., & Bradley, M. M. (1998). Measuring emotion in therapy: Imagery, activation, and feeling. *Behavior Therapy*, *29*(4), 655-674.

[54] Shah, A., Jhawar, S. S., & Goel, A. (2012). Analysis of the anatomy of the Papez circuit and adjoining limbic system by fiber dissection techniques. *Journal of Clinical Neuroscience*, 19(2), 289-298.

[55] Hahn, T., Dresler, T., Plichta, M. M., Ehlis, A. C., Ernst, L. H., Markulin, F., ... & Fallgatter, A. J. (2010). Functional amygdala-hippocampus connectivity during anticipation of aversive events is associated with Gray's trait "sensitivity to punishment". *Biological psychiatry*, *68*(5), 459-464.

[56] Corr, P. J. (2013). Approach and avoidance behaviour: Multiple systems and their interactions. *Emotion Review*, *5*(3), 285-290.

[57] Gable, P. A., Mechin, N., Hicks, J., & Adams, D. L. (2015). Supervisory control system and frontal asymmetry: neurophysiological traits of emotion-based impulsivity. *Social cognitive and affective neuroscience*, nsv017.

[58] DiMaggio, P. (1997). Culture and cognition. Annual review of sociology, 263-287.

[59] Miller, E. K. (2000). The prefontral cortex and cognitive control. *Nature reviews neuroscience*, *1*(1), 59-65.

[60] Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, *24*(1), 167-202.

[61] Solomon, M., Ozonoff, S. J., Ursu, S., Ravizza, S., Cummings, N., Ly, S., & Carter, C. S. (2009). The neural substrates of cognitive control deficits in autism spectrum disorders. *Neuropsychologia*, *47*(12), 2515-2526.

[62] Tupak, S. V., Dresler, T., Badewien, M., Hahn, T., Ernst, L. H., Herrmann, M. J., ... & Fallgatter, A. J. (2013). Inhibitory transcranial magnetic theta burst stimulation attenuates prefrontal cortex oxygenation. Human brain mapping, 34(1), 150-157.

[63] Miller, B. T., & D'Esposito, M. (2005). Searching for "the top" in top-down control. *Neuron*, *48*(4), 535-538.

[64] Dias, A. R. N. (2014). *Mechanistic characterization of reinforcement learning in healthy humans using computational models*. (Master's Thesis, Instituto Superior Técnico, Lisbon, Portugal).

[65] Thorndike, E. L. (1911). *Animal intelligence: Experimental studies*. Macmillan.

[66] Bayley, P. J., Frascino, J. C., & Squire, L. R. (2005). Robust habit learning in the absence of awareness and independent of the medial temporal lobe. *Nature*, *436*(7050), 550-553.

[67] Daw, N. D., Niv, Y., & Dayan, P. (2005). Recent breakthroughs in basal ganglia research. *chap. Actions, policies, values, and the basal ganglia). Nova science publishers*, *113*.

[68] Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, *7*(6), 464-476.

[69] Neal, D. T., Wood, W., & Quinn, J. M. (2006). Habits—A repeat performance. Current Directions in Psychological Science, 15(4), 198-202.

[70] Gasbarri, A., Pompili, A., Packard, M. G., & Tomaz, C. (2014). Habit learning and memory in mammals: Behavioral and neural characteristics. Neurobiology of learning and memory, 114, 198-208.

[71] Lally, P., Van Jaarsveld, C. H., Potts, H. W., & Wardle, J. (2010). How are habits formed: Modelling habit formation in the real world. European Journal of Social Psychology, 40(6), 998-1009.

[72] Aarts, H., & Dijksterhuis, A. (2000). Habits as knowledge structures: automaticity in goal-directed behavior. Journal of personality and social psychology, 78(1), 53.

[73] Jones, C. R., Vilensky, M. R., Vasey, M. W., & Fazio, R. H. (2013). Approach behavior can mitigate predominately univalent negative attitudes: Evidence regarding insects and spiders. Emotion, 13(5), 989.

[74] Cretenet, J., & Dru, V. (2004). The influence of unilateral and bilateral arm flexion versus extension on judgments: an exploratory case of motor congruence. Emotion, 4(3), 282.

[75] Centerbar, D. B., & Clore, G. L. (2006). Do approach-avoidance actions create attitudes?. Psychological science, 17(1), 22-29.

[76] Kawakami, K., Phills, C. E., Steele, J. R., & Dovidio, J. F. (2007). (Close) distance makes the heart grow fonder: Improving implicit racial attitudes and interracial interactions through approach behaviors. Journal of personality and social psychology, 92(6), 957.

[77] Huijding, J., Muris, P., Lester, K. J., Field, A. P., & Joosse, G. (2011). Training children to approach or avoid novel animals: Effects on self-reported attitudes and fear beliefs and information-seeking behaviors. Behaviour research and therapy, 49(10), 606-613.

[78] Rinck, M., & Becker, E. S. (2007). Approach and avoidance in fear of spiders. *Journal of behavior therapy and experimental psychiatry*, *38*(2), 105-120.

[79] Phaf, R. H., Mohr, S. E., Rotteveel, M., & Wicherts, J. M. (2014). Approach, avoidance, and affect: a meta-analysis of approach-avoidance tendencies in manual reaction time tasks. *Frontiers in psychology*, *5*.

[80] Wiers, C. E., Stelzel, C., Park, S. Q., Gawron, C. K., Ludwig, V. U., Gutwinski, S., ... & Bermpohl, F. (2014). Neural correlates of alcohol-approach bias in alcohol addiction: the spirit is willing but the flesh is weak for spirits. *Neuropsychopharmacology*, *39*(3), 688-697.

[81] Volman, I., Toni, I., Verhagen, L., & Roelofs, K. (2011). Endogenous testosterone modulates prefrontal–amygdala connectivity during social emotional behavior. *Cerebral Cortex*, bhr001.

[82] Roelofs, K., Minelli, A., Mars, R. B., van Peer, J., & Toni, I. (2009). On the neural control of social emotional behavior. *Social Cognitive and Affective Neuroscience*, *4*(1), 50-58.

[83] Wiers, R. W., Rinck, M., Dictus, M., & Van den Wildenberg, E. (2009). Relatively strong automatic appetitive action-tendencies in male carriers of the OPRM1 G-allele. *Genes, Brain and Behavior*, *8*(1), 101-106.

[84] http://mscoco.org/dataset/#download Microsoft Common Objects in Context accessed in 10/06/2015.

[85] Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). International affective picture system (IAPS): Affective ratings of pictures and instruction manual.*Technical report A-8*.

[86] Ratcliff, R. (1993). Methods for dealing with reaction time outliers.*Psychological bulletin*, *114*(3), 510.

[87] Miller, J. (1991). Short report: Reaction time analysis with outlier exclusion: Bias varies with sample size. *The quarterly journal of experimental psychology*,*43*(4), 907-912.

[88] Whelan, R. (2010). Effective analysis of reaction time data. *The Psychological Record*, *58*(3), 9.

[89] Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic bulletin & review*, *7*(3), 424-465.

[90] Bates, D. M. (2010). lme4: Mixed-effects modeling with R. *URL http://lme4. r-forge. r-project. org/book*.

[91] Pinheiro, J., & Bates, D. (2006). *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media.

[92] Krueger, C., & Tian, L. (2004). A comparison of the general linear mixed model and repeated measures ANOVA using a dataset with multiple missing data points. *Biological research for nursing*, *6*(2), 151-157.

[93] Quené, H., & Van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*(4), 413-425.

[94] Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255-278.

[95] Neal, D. T., Wood, W., & Quinn, J. M. (2006). Habits—A repeat performance.*Current Directions in Psychological Science*, *15*(4), 198-202.

[96] Eder, A. B. (2011). Control of impulsive emotional behaviour through implementation intentions. *Cognition and Emotion*, *25*(3), 478-489.

[97] Wiers, R. W., Gladwin, T. E., Hofmann, W., Salemink, E., & Ridderinkhof, K. R. (2013). Cognitive bias modification and cognitive control training in addiction and related psychopathology mechanisms, clinical perspectives, and ways forward. *Clinical Psychological Science*, 2167702612466547.

[98] Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *The Journal of Neuroscience*, *32*(2), 551-562.

[99] Collins, A. G., & Frank, M. J. (2014). Opponent actor learning (OpAL): Modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychological review*, *121*(3), 337.

[100] Huys, Q. J., Cools, R., Gölzer, M., Friedel, E., Heinz, A., Dolan, R. J., & Dayan, P. (2011). Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. *PLoS Comput Biol*, *7*(4), e1002028.

[101] McCall, C., Tipper, C. M., Blascovich, J., & Grafton, S. T. (2012). Attitudes trigger motor behavior through conditioned associations: neural and behavioral evidence. *Social cognitive and affective neuroscience*, *7*(7), 841-849.

[102] Pins, D., & Bonnet, C. (1996). On the relation between stimulus intensity and processing time: Piéron's law and choice reaction time. *Perception & psychophysics*, *58*(3), 390-400.

[103] Van Maanen, L., Grasman, R. P. P. P., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Piéron's Law and Optimal Behavior in Perceptual Decision-Making. *Frontiers in Neuroscience*, *5*, 143.

[104] Stafford, T., Ingram, L., & Gurney, K. N. (2011). Piéron's law holds during Stroop conflict: insights into the architecture of decision making. *Cognitive science*, *35*(8), 1553-1566.

[105] Stan Development Team. 2015. *Stan Modeling Language Users Guide and Reference Manual, Version 2.8.0.*

[106] Daw, N. D. (2011). Trial-by-trial data analysis using computational models. *Decision making, affect, and learning: Attention and performance XXIII*, *23*, 1.

[107] Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. University of London.

[108] Bo, J., Jennett, S., & Seidler, R. D. (2011). Working memory capacity correlates with implicit serial reaction time task performance. *Experimental brain research*, *214*(1), 73-81.

[109] Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic bulletin & review*,*7*(2), 185-207.

[110] Kutner, M. H. (1996). Applied linear statistical models (Vol. 4). Chicago: Irwin.

[111] http://www.itl.nist.gov/div898/handbook/eda/section3.htm Engineering Statistics Handbook accessed in 6/8/2015.

[112] Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *Neuroimage*, *46*(4), 1004-1017.

[113] Penny, W. D., Stephan, K. E., Mechelli, A., & Friston, K. J. (2004). Comparing dynamic causal models. *NeuroImage*, *22*(3), 1157-1172.

[114] Penny, W. D. (2012). Comparing dynamic causal models using AIC, BIC and free energy. *Neuroimage*, *59*(1), 319-330.

[115] Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies—Revisited. *NeuroImage*, *84*, 971-985.

[116] Daunizeau, J., Adam, V., & Rigoux, L. (2014). VBA: a probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLoS computational biology*, *10*(1).

[117] Asmussen, S., & Rojas-Nandayapa, L. (2008). Asymptotics of sums of lognormal random variables with Gaussian copula. *Statistics & Probability Letters*, *78*(16), 2709-2714.

[118] Eberl, C., Wiers, R. W., Pawelczack, S., Rinck, M., Becker, E. S., & Lindenmeyer, J. (2014). Implementation of Approach Bias Re-Training in Alcoholism—How Many Sessions are Needed?. *Alcoholism: Clinical and Experimental Research*, *38*(2), 587-594.

[119] Shen, W. H. (1998). Estimation of parameters of a lognormal distribution. *Taiwanese Journal of Mathematics*, *2*(2), pp-243.

[120] Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1997). Beyond bipolar conceptualizations and measures: The case of attitudes and evaluative space.*Personality and Social Psychology Review*, *1*(1), 3-25.

[121] Peduzzi, P., Concato, J., Feinstein, A. R., & Holford, T. R. (1995). Importance of events per independent variable in proportional hazards regression analysis II. Accuracy and precision of regression estimates. Journal of clinical epidemiology, 48(12), 1503-1510.

[122] Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods?. *American Psychologist*, *53*(3), 300.

[123] Razali, N. M., & Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics*, *2*(1), 21-33.

A

Appendices

Contents

A.1 Task implementation features

In order to perform a specific version of the task and set the different features of each version, the experimenter was allowed to fill in a specific excel file, whose structure was already predefined (cf. figure A1.1). Therefore 4 excel files were created: Train_test.xlsx, Train_NN.xlsx, Train_PN.xlsx, and Test.xlsx. The first one was used for the arrow version, the following two were used when performing the train for the negative and positive group, respectively, and the last one was used for the Assessment version.

One example of this structure is provided in the figure A1.1 as well as a brief explanation of each column.

Picture Category	Instructed Action	Nr Trials/	Folder	Output Folder
Neutral	Approach	6	C:\Users\User\Dropbox\FMUL - Investigação\Joystick Task\stimuli\NeutralApproach_test\	C:\Users\User\Dropbox\FMUL - Investigação\Joystick Task\Outputs
Neutral	Avoid	6	C:\Users\User\Dropbox\FMUL - Investigação\Joystick Task\stimuli\NeutralAvoid_test\	
Negative	Approach	6	C:\Users\User\Dropbox\FMUL - Investigação\Joystick Task\stimuli\NegativeApproach_test\	
Positive	Avoid	6	C:\Users\User\Dropbox\FMUL - Investigação\Joystick Task\stimuli\PositiveAvoid_test\	

Figure A1.1: Structure of the excel file Test.xlsx.

Picture Category: column where the category of the stimuli is specified.

Instructed Action: column that represents the action that the subject was instructed to perform.

Nr. Trials/ pic.: column where the experimenter sets the number of repetitions of each picture.

Folder: column that refers to the directory where the MATLAB routine selects the pictures for the conditions, specified in the first two columns.

Output Folder: column where the experimenter sets the directory to where the output file generated for each subject shall be stored.

Output data

For each subject four output files were created and the data were saved in the file “Outputs” as “control_(version)_(subject number).(extension)” (e.g.: *control_train_1.xlsx*; *control_train_1.csv*; *control_trainEva_1.xlsx*; *control_trainEva_1.csv*). This was done to facilitate the output readout by the different softwares used for the analysis.

The file only had one sheet and the identification of the group each subject belonged was done in an auxiliary Excel file. The latter was also used to randomly distribute the subjects for groups and to randomly assign the pictures.

Regarding the output files generated, in the case of the first two examples above, the data was organized in a table with 8 columns, whose rows correspond to each trial, and we described next.

Trial_Nr.: The trial number.

Picture_Category: Categorical variable which codes the picture category: positive (1), negative (2) and neutral (3).

Picture_Nr.: The number of the picture presented in a trial.

Instructed_Action: Categorical variable which codes whether the subject was instructed to pull, *i.e.*, approach (1) or push, *i.e.*, avoid (2) the joystick.

Action_Chosen: Categorical variable which codes whether the subject chose to pull, *i.e.*, approach (1) or push, *i.e.*, avoid (2) the joystick.

First_Movement: Categorical variable that codes if the first movement of the subject was accordingly to what was instructed (0) or against it (1).

RT_until_movement: Time the subject take to make the first movement with the joystick, since the stimulus presentation. It is measured in seconds.

Movement_RT: Time the subject take to complete the trial after the first movement. It is measured in seconds.

In the case of the last two examples of the output files, the data was organized in a table with 3 columns, whose rows correspond to the number of the picture. The description is given below.

First: Represents the rating of a certain stimulus selected by the subject.

Instruction: Categorical variable which codes whether the subject was instructed to pull, *i.e.*, approach (1) or push, *i.e.*, avoid (2) the joystick.

Category: Categorical variable which codes the picture category: positive (1), negative (2) and neutral (3).

Running the routine

When the user runs the task, a dialog box is displayed for the user to select the input file corresponding to the version of the task the subject must perform (cf. figure A1.2a). Then the input file is immediately read and a new dialog box shows up with three blank boxes (cf. figure A1.2b) where the user must insert the number of the subject, the number of the session and the number associated to the version that will be performed.

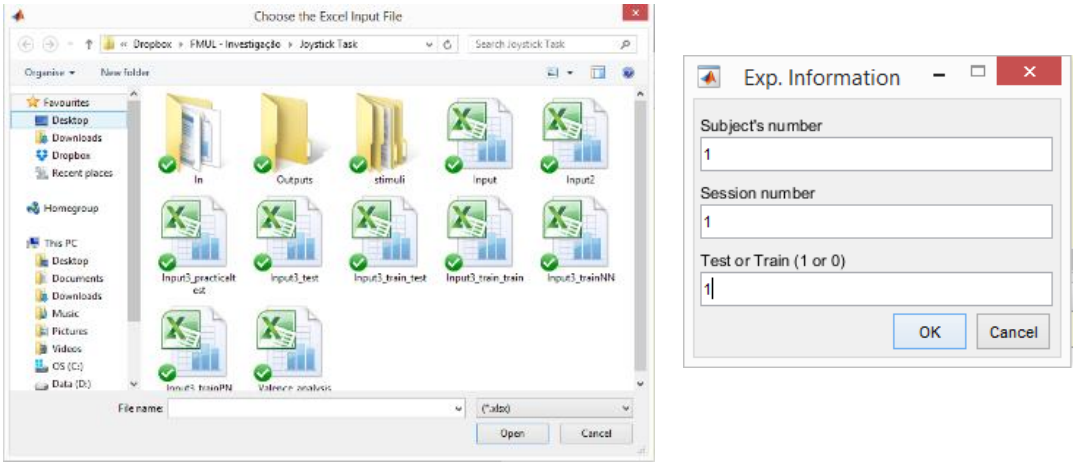


Figure A1.2: Dialogue box displayed at the beginning of the task (a) to select the excel input file and (b) to fill in the blank boxes where the experimenter must insert the subject's number, the number of session and number associated to the routine that will be performed.

Then the settings to perform the task are settled and a function we created named *Load_images_Trials_creation* loads the images necessary to the execution of the routine as well as creates a matrix with the trials the subject will perform. For statistical reasons, after the trial matrix is created, we proceed to a permutation of the several conditions it contains.

A.2 Results from the first sample

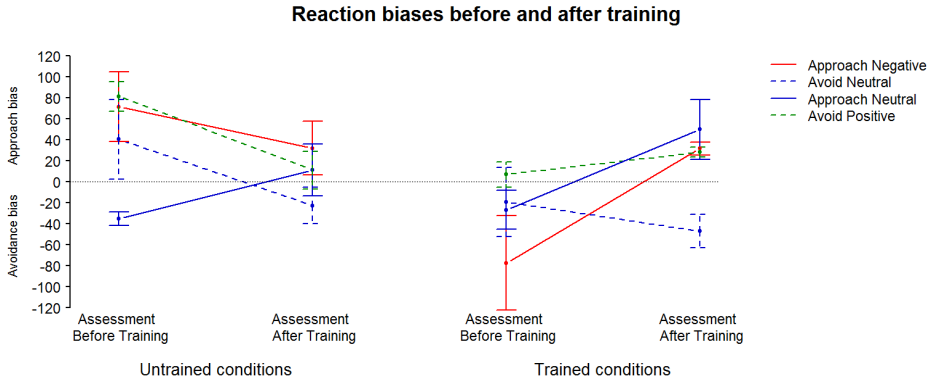


Figure A2.1: Results from the assessment version of the AAT that depict the reaction biases before and after the training of the positive group, towards negative and neutral pictures they did not train, and for the negative group, towards the positive and neutral pictures they did not train (left side), and reaction biases before and after the training of positive group, towards the conditions they trained, and for the negative group, towards the conditions they trained (right side).

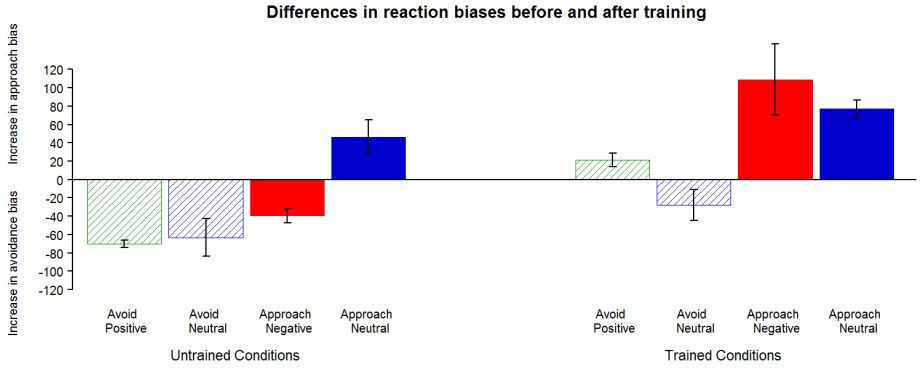


Figure A2.2: Results from the assessment version of the AAT. The differences in reaction biases before and after training of the positive group, towards negative and neutral pictures they did not train, and for the negative group, towards the positive and neutral pictures they did not train (left side), and the differences in reaction biases before and after training of positive group, towards the conditions they trained, and for the negative group, towards the conditions they trained (right side).

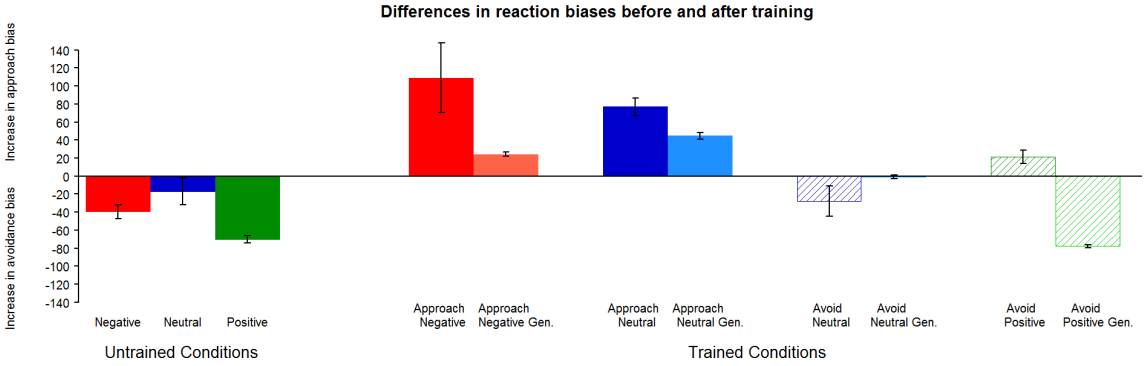


Figure A2.3: Results from the assessment version of the AAT. The differences in reaction biases before and after training of the positive group, towards negative and neutral pictures they did not train, and for the negative group, towards the positive and neutral pictures they did not train (the three initial bars) and the differences in reaction biases before and after training of positive group, towards the conditions they trained, and for the negative group, towards the conditions they trained, including results obtained for generalization, respectively (right side).

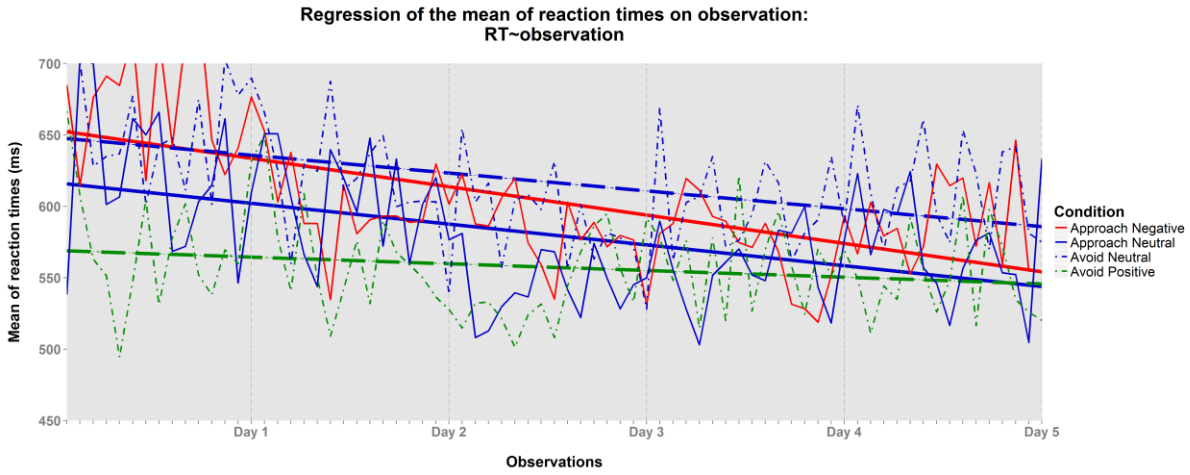


Figure A2.4: Results of the linear fit performed to each condition using the mean of the reaction times between pictures of that condition and between the subjects that trained that condition.

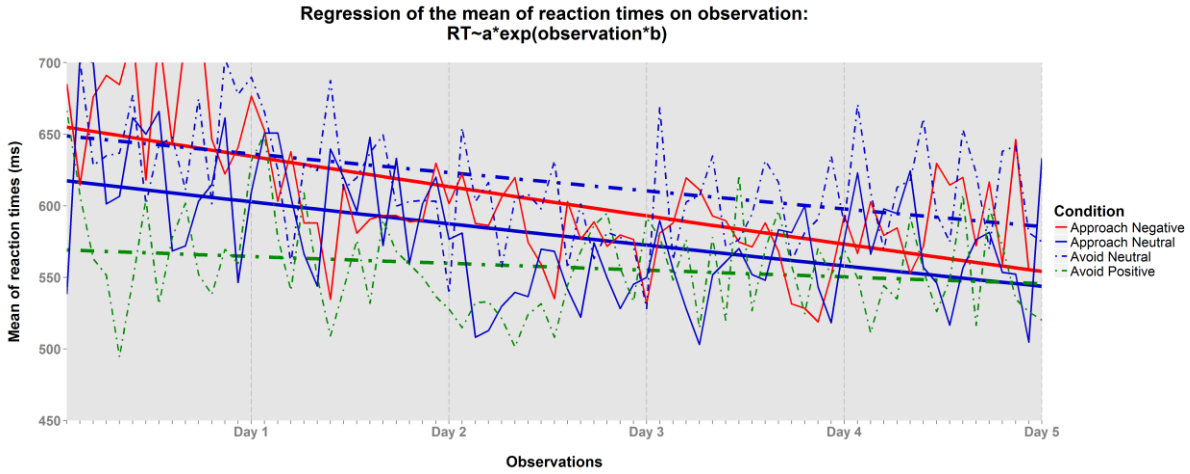


Figure A2.5: Results of the exponential fit performed to each condition using the mean of the reaction times between pictures of that condition and between the subjects that trained that condition.

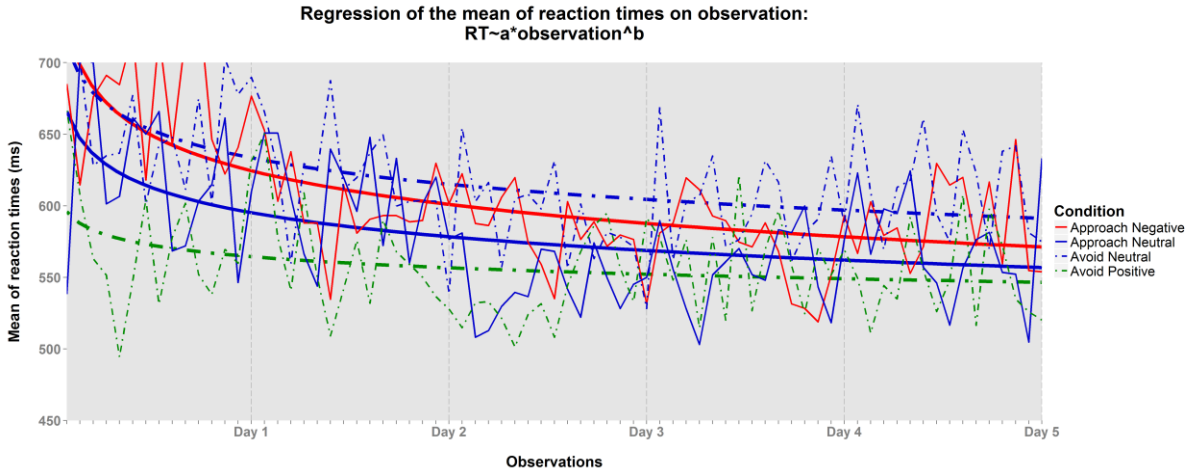


Figure A2.6: Results of the power law fit performed to each condition using the mean of the reaction times between pictures of that condition and between the subjects that trained that condition.

A.3 Complement of the exploratory data analysis

Subject	Group	Distribution	$p-value$ of the Kolmogorov-Smirnov test					
			Day 1	Day 2	Day 3	Day 4	Day 5	All days
Subject 1	Negative	Normal	0.398	0.654	0.047	0.282	0.030	0.023
		Log-Normal	0.291	0.554	0.157	0.489	0.158	0.070
Subject 2	Negative	Normal	0.071	0.157	0.240	0.835	0.049	<0.001
		Log-Normal	0.226	0.423	0.518	0.578	0.151	0.002
Subject 3	Negative	Normal	0.165	0.237	0.003	0.053	0.233	<0.001
		Log-Normal	0.547	0.773	0.033	0.295	0.489	0.011
Subject 4	Negative	Normal	0.057	0.151	0.114	0.164	0.015	<0.001
		Log-Normal	0.414	0.558	0.215	0.623	0.209	0.003
Subject 5	Negative	Normal	0.003	0.081	0.003	0.015	0.005	<0.001
		Log-Normal	0.054	0.363	0.039	0.184	0.150	<0.001
Subject 6	Negative	Normal	0.039	0.145	0.242	0.368	0.274	<0.001
		Log-Normal	0.131	0.290	0.432	0.267	0.592	0.009
Subject 7	Negative	Normal	0.179	0.040	0.302	0.089	0.100	<0.001
		Log-Normal	0.538	0.058	0.498	0.176	0.319	0.003
Subject 8	Negative	Normal	0.027	0.109	0.033	0.062	0.145	<0.001
		Log-Normal	0.285	0.456	0.087	0.216	0.595	0.017
Subject 9	Negative	Normal	0.547	0.106	0.391	0.181	0.711	0.001
		Log-Normal	0.822	0.201	0.879	0.688	0.763	0.270
Subject 10	Negative	Normal	0.122	0.237	0.840	0.125	0.654	0.005
		Log-Normal	0.337	0.329	0.435	0.440	0.891	0.362
Subject 11	Negative	Normal	0.332	0.472	0.246	0.679	0.851	0.022
		Log-Normal	0.401	0.827	0.357	0.486	0.960	0.388
Subject 12	Negative	Normal	0.208	0.503	0.511	0.513	0.057	<0.001
		Log-Normal	0.065	0.207	0.466	0.979	0.274	0.028
Subject 13	Negative	Normal	0.124	0.431	0.144	0.551	0.845	0.016
		Log-Normal	0.174	0.386	0.346	0.951	0.950	0.022
Subject 14	Negative	Normal	0.468	0.082	0.920	0.627	0.436	0.009
		Log-Normal	0.860	0.295	0.548	0.241	0.548	0.448
Subject 15	Negative	Normal	0.291	0.388	0.524	0.451	0.320	0.007
		Log-Normal	0.706	0.659	0.659	0.699	0.718	0.348
Subject 16	Negative	Normal	0.300	0.621	0.749	0.909	0.633	0.002
		Log-Normal	0.841	0.638	0.993	0.659	0.431	0.031
Subject 17	Negative	Normal	0.337	0.792	0.357	0.411	0.802	0.002
		Log-Normal	0.710	0.903	0.832	0.791	0.961	0.015
Subject 18	Negative	Normal	0.519	0.430	0.696	0.399	0.460	0.002
		Log-Normal	0.465	0.789	0.977	0.757	0.806	0.108
Subject 1	Positive	Normal	0.478	0.025	0.318	0.090	0.085	<0.001
		Log-Normal	0.577	0.131	0.843	0.460	0.501	0.022
Subject 2	Positive	Normal	0.231	0.252	0.088	0.116	0.302	0.040
		Log-Normal	0.558	0.640	0.276	0.264	0.595	0.018
Subject 3	Positive	Normal	0.200	0.023	0.064	0.601	0.528	0.044
		Log-Normal	0.518	0.198	0.344	0.807	0.171	0.010
Subject 4	Positive	Normal	0.068	0.238	0.001	0.009	0.070	<0.001
		Log-Normal	0.220	0.646	0.017	0.089	0.421	<0.001
Subject 5	Positive	Normal	0.471	0.173	0.413	0.319	0.066	<0.001
		Log-Normal	0.921	0.559	0.664	0.719	0.300	0.028
Subject 6	Positive	Normal	0.027	0.032	0.234	0.012	0.066	<0.001
		Log-Normal	0.217	0.162	0.414	0.060	0.165	<0.001
Subject 7	Positive	Normal	0.434	0.127	0.282	0.064	0.100	<0.001
		Log-Normal	0.352	0.284	0.529	0.224	0.139	0.012
Subject 8	Positive	Normal	0.181	0.197	0.423	0.294	0.078	0.038
		Log-Normal	0.500	0.500	0.696	0.443	0.810	0.045
Subject 9	Positive	Normal	0.061	0.417	0.055	0.121	0.142	<0.001
		Log-Normal	0.186	0.237	0.101	0.447	0.264	<0.001
Subject 10	Positive	Normal	0.006	0.097	0.129	0.192	0.012	0.000
		Log-Normal	0.019	0.448	0.406	0.649	0.132	0.003
Subject 11	Positive	Normal	0.299	0.003	0.019	0.276	0.081	<0.001
		Log-Normal	0.514	0.015	0.103	0.565	0.182	<0.001
Subject 12	Positive	Normal	0.132	0.584	0.001	0.056	0.002	<0.001
		Log-Normal	0.515	0.658	0.022	0.232	0.074	<0.001
Subject 13	Positive	Normal	0.509	0.093	0.364	0.102	0.317	<0.001
		Log-Normal	0.471	0.182	0.445	0.255	0.759	0.006
Subject 14	Positive	Normal	0.048	0.177	0.180	0.305	0.080	<0.001
		Log-Normal	0.349	0.430	0.592	0.642	0.347	0.002
Subject 15	Positive	Normal	0.348	0.729	0.819	0.978	0.061	0.048
		Log-Normal	0.919	0.525	0.720	0.932	0.242	0.036
Subject 16	Positive	Normal	0.403	0.689	0.837	0.243	0.821	0.031
		Log-Normal	0.728	0.548	0.974	0.277	0.770	0.222
Subject 17	Positive	Normal	0.532	0.185	0.533	0.809	0.067	<0.001
		Log-Normal	0.888	0.456	0.593	0.830	0.182	0.003
Subject 18	Positive	Normal	0.563	0.251	0.716	0.571	0.432	0.002
		Log-Normal	0.908	0.626	0.916	0.975	0.936	0.022

Table A3.1: Results of the *Kolmogorov-Smirnov* test applied to each session (day) of all subjects.

A.4 Individual results from the arrow version of the AAT

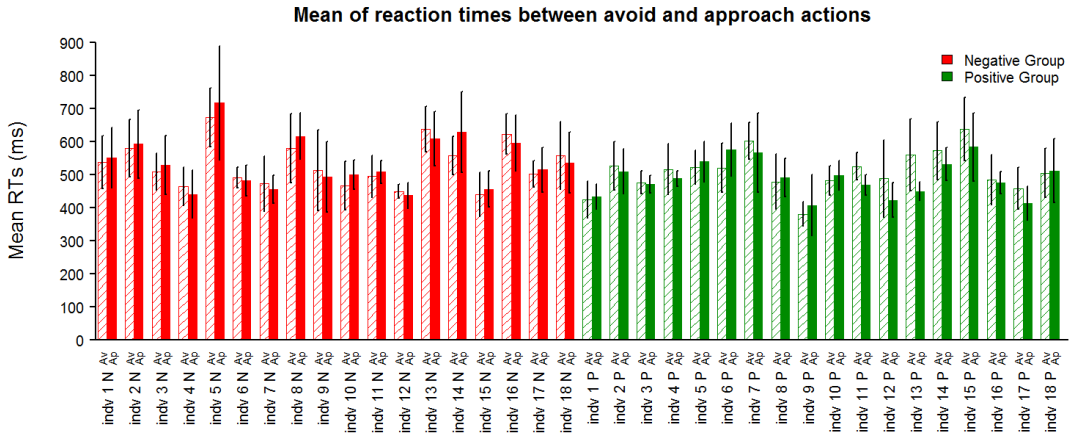


Figure A4.1 Individual RTs acquired by the arrow version of the AAT.

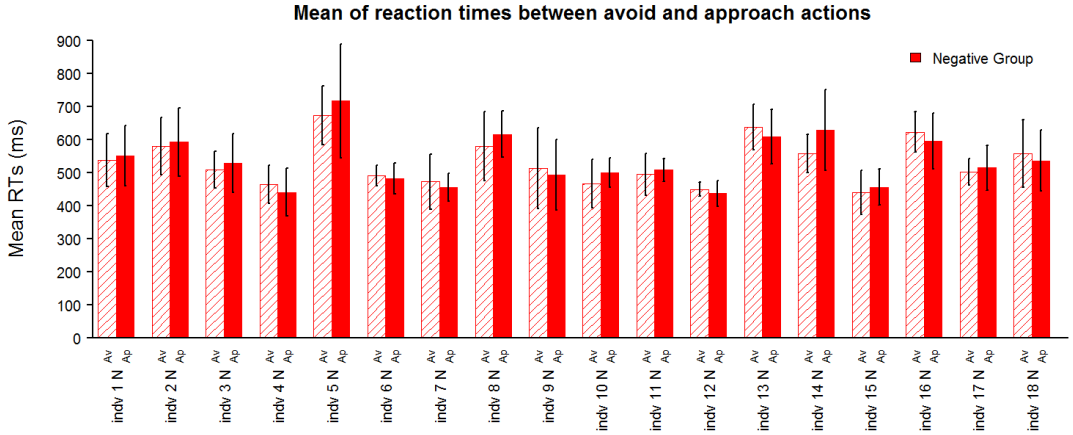


Figure A4.2: Individual RTs acquired by the arrow version of the AAT for the negative group.

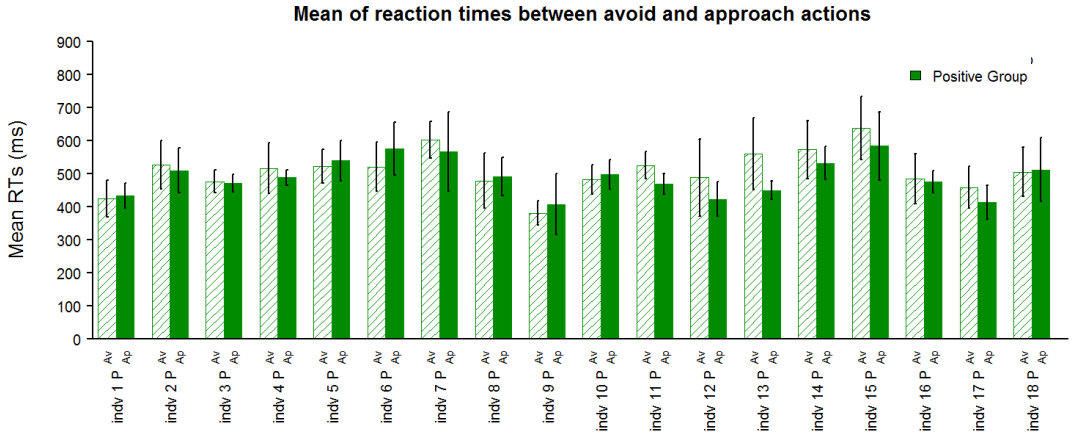


Figure A4.3: Individual RTs acquired by the arrow version of the AAT for the positive group.

A.5 BMC between the first four models with the transformed RTs through the moving average method

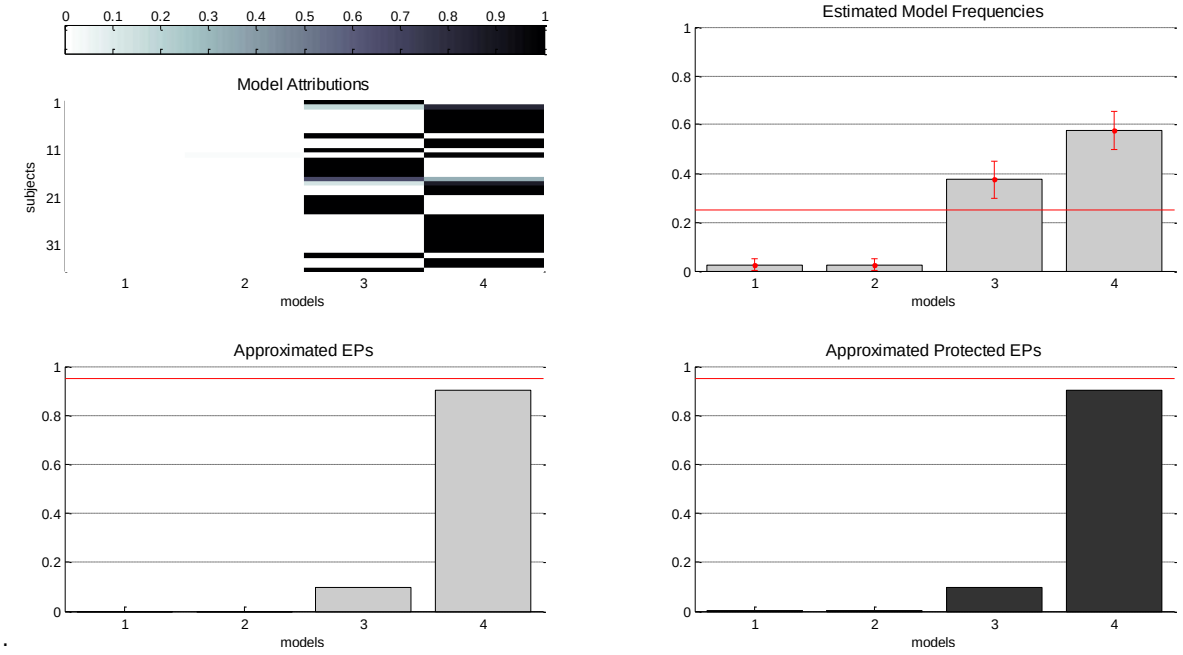


Figure A5.1: Results of BMC between the first four models which received as input transformed RTs (via moving average filtering) described in table 3.2 obtained using the BIC, for the population. (Top Left) Model Attribution at subject level, the first eighteen subjects belong to negative group, while the last eighteen to the positive group. (Top Right) Estimated model frequencies. (Left and Right Bottom) Approximated Exceedance Probabilities (EPs) and Protected EPs of the eight models.

As it is visible in figure A5.1, when we only compare the first four models the outcome of the BMC showed a trend since a PEP of 90% was obtained for the fourth model. Besides this clearly shows there is a considerable difference between using the Normal likelihood function and the Log-Normal likelihood function.