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

Tiago Henriques
Instituto Superior Técnico
Instituto de Medicina Molecular
Faculdade de Medicina da Universidade de Lisboa

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.

Key words

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

1. Introduction

OCD patients have predispositions toward excessive stereotyped behavior in order to avoid adverse consequences [1]. Therefore, it is assumed that some OCD symptoms might be seen as enhanced avoidance tendencies [2] [3] [4].

However, the standard treatment, Exposure and Response Prevention (ExRP) is substantially discomforting, time-consuming and associated with a high relapse rate [5] [6] [7].

Consequently, it is of utmost importance to find ways to make this treatment less aversive. Thereby, we first have to understand how people, in general, regulate their automatic tendencies and engage in so-called controlled responses that may be against their primary reaction tendencies, but are advantageous for the achievement of long term-goals [8].

So, to disentangle the interplay between impulsive automatic reactions and their inhibition through cognitive control and to identify the cognitive subcomponents involved, we resorted to the AAT.

The AAT is an extensively used implicit task that directly assesses the behavioral component of approach-avoidance impulses and also allows to evaluate the deliberative regulation of these impulses [8]. In fact, the AAT was proven useful to understand certain aspects of psychopathology, such as the pathologically enhanced approach tendencies for alcohol-related stimuli in alcohol dependence [9][10] and for heroin-related stimuli in people addicted to heroin [11], and the pathologically enhanced avoidance tendencies for phobia-relevant stimuli in different phobias [12].

Moreover, training with the AAT has been also associated to findings that show evidence for a behavior modification transduced in the overcoming of the automatic action tendencies, through repetition of incongruent reactions [10] [13] [14] [15] [16] [17]. Consequently, modifications of the stimuli's valence processing have been considered as consequence of this training [10] [13] [15] [17] [18] [19].

Neuronal correlates: automatic and regulated processes

The definition of approaching positive stimuli and avoiding negative ones as vital behaviors was based on the fact that there are specialized nervous systems to processes them.

In fact, it was demonstrated that defensive reflexes in humans might rely on the same neuronal structures as does the fear circuit, specifically on the amygdala [20] [21]. On the other hand, further investigations linked the activity of the Basal Ganglia (BG) to approach behavior [22]. Then, the regulation and control of such behaviors has been associated to the Pre-Frontal-Cortex (PFC) [23] [21] [24]. Moreover, electrophysiological studies showed evidence for specific approach-avoidance systems regarding hemispheric asymmetry [21] [25] [26].

Besides, 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) and the ones that are controlled and follow a plan to overcome these automatic reactions [27].

In fact, this assumption led to the development of the Reflective Impulse Model [28]. This model states that the impulsive system is engaged through perceptual input, therefore 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. However, the latter is highly dependent on the allocation of control and attentional resources [28] [29] [30] [31].

Instrumental conditioning

Instrumental conditioning is one type of learning in which the subject's behavior is adjusted accordingly to its consequences [32]. This complex form of behavior was formally postulated in Thorndike's Law of effect which states that responses followed by a positive (negative) outcome would strengthen (weaken) the stimulus-response (S-R) association, *i.e.*, the formation of a new habit. Besides, instrumental learning also takes into account the fact that learning occurs via stimulus-response-outcome (S-R-O) associations, the so-called goal-directed behavior [33] [34] [35] [36].

Although behavior might be ruled by a goal-directed approach at first, 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 this action.

Indeed, this process that was termed "habit formation" was shown to be very important, since much of our everyday life action is steered by repetition [37]: Habit formation results from the acquisition of sequential and repetitive motor behaviors triggered by external or internal stimuli. This results in incremental strengthening of the S-R association leading to an increase of the automaticity of such behaviors and consequently reduces the cognitive load necessary to perform such actions. [34] [38] [39] [40].

However, neither the aforementioned cognitive processes nor the learning processes that are influenced and involved by the training version of the AAT were investigated thus far.

There is, therefore, an obvious need of better understanding the training version of the AAT considering the psychological and cognitive processes associated to the performance of incongruent reactions, such as approaching the negative and avoiding positive stimuli. In fact, recent findings of neuronal substrates involved in the performance of these reactions [10] [11] [13] [15] [27] and the substantial body of work, highlighting the usefulness of computational reinforcement-learning models in characterizing behavior and brain-behavior relationships, [33] [41] [42] led to the design of a novel computational framework. This will then allow us to characterize relevant aspects of the cognitive processes involved in the referred version of the AAT.

While there is still a long way to go before a completely reliable system for the quantification and capture of these processes can be developed, these computational approaches constitute, therefore, an important and promising tool to fulfill this goal.

2. Methods

In order to assess different behavioral aspects that underlie the overcoming of automatic response tendencies when performing the AAT, 36 healthy subjects (divided in two balanced groups, the negative and positive groups) were assessed through three different versions, created through modification of the original version of the joystick version with feedback. The task versions next described were developed in MATLAB version R2012b, using Psychophysics Toolbox Version 3 (PTB-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. Besides, all participants went through a small train to perceive the joystick sensibility and the instructions that they would receive. 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.

All participants followed a protocol of 5 consecutive days, in which 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.

Training version of the AAT

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. This routine was performed during 5 consecutive days.

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 1.

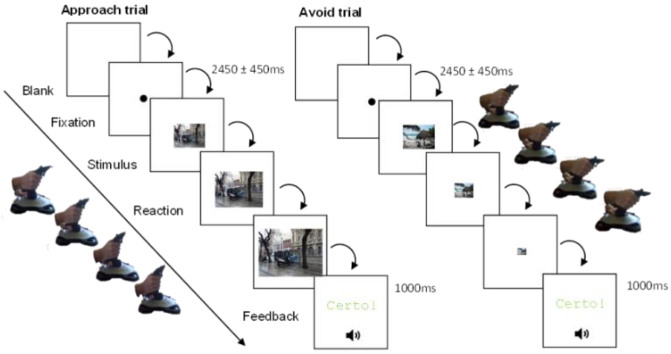
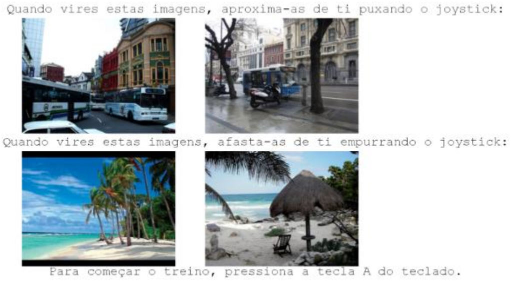


Figure 1: Schematics of a typical trial in the task composed of 5 events: fixation, stimulus, reaction, feedback and blank. Temporal differences between each event are also depicted for both types of trials. The sonorous icon corresponds to the sound that accompanied the visual feedback.

After completing the block of 120 trials, we asked participants to rate the pictures they were presented with, in terms of pleasantness. Besides the verbal instructions provided we also instructed participants according to figure 2.



Quando vires estas imagens, aproxima-as de ti puxando o joystick:

Quando vires estas imagens, afasta-as de ti empurrando o joystick:

Para começar o treino, pressiona a tecla A do teclado.

Figure 2: 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".

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 [13].

Assessment version of the AAT

The assessment version was created in order to capture the effects of the unintentional valence processing of participants and to verify whether participants generalized the learning to similar pictures. In this version, which consisted of a routine of 192 trials in total, we presented 16 pictures to participants that were grouped in 3 different categories (cf Stimuli). 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 figure 1, except for some modifications. Firstly, the stimuli presentation was done along with the instruction of which action the participant should perform (figures 3a and 3b). Secondly, participants only received feedback when they wrongly performed one trial. Thirdly, a written instruction was provided 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 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 [43].

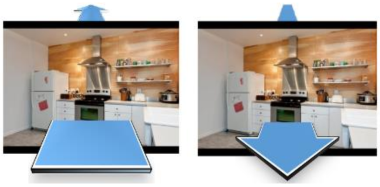


Figure 3: General instructions and stimuli provided to participants in the assessment version. (a) Indicates that the participant should avoid the picture by pushing the joystick and (b) indicates that the participant should approach the picture by pulling the joystick.

Arrow version of the AAT

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. This version was designed in order to assess participants’ motor biases when using the joystick. Therefore we used a similar trial structure to the one of figure 1, where the stimuli presentation and instruction provided were similar to figure 3a and 3b, but instead of one image we used a black rectangle.

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. 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 ones depicted in figure 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 (negative: 154718, 264964; neutral: 641, 10114, 10844, 18366, 20979, 26204, 27285, 28682; positive: 348896, 490337, 556420) [44], from the database of the IAPS (negative: 9300, 9320) [45] and from internet (positive: ‘c30’).

Behavioral dataset

The 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 standard deviation (SD) = 1.754 years. Due to hardware limitations (just one joystick and one computer) the acquisition of data was done during three weeks and the participants’ data were collected according to their availability. 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. 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. All the participants were acquaintances of the experimenter and performed the tasks voluntarily. All provided written informed consent. The study was approved by the local Ethics Committee for the Health Care of the University of Lisbon.

Data pre-processing and outliers’ exclusion criteria

Firstly, according to prior analyses of reaction times acquired during the solving of behavioral tasks [10] [13] [14] [46], 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.

Then, we also analyzed the data for additional outliers and considering the literature related to the analysis of RTs, several criteria for pre-processing of correct RTs were found [47] [48] [49]. We decided to use a specific cut-off at 200 ms, because we did not expect meaningful RTs to be faster than this for our task and the cut-off at 3 times the interquartile range above the third quartile, which prevented to eliminate meaningful information as might happen with the previously used cut-off at 3 standard deviations above the mean. These criteria and analysis were performed at individual level.

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 [50]. 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 [51]. 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 1:

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{b} + \boldsymbol{\varepsilon} \qquad (1)$$

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 $\boldsymbol{y}$.

Besides the use of these models to fit the behavioral data acquired from the training, where random effects were considered for both the intercept and the slope, we also used them to analyze the behavioral data acquired from the other AAT routines instead of the classical ANOVAS. This because mixed models were shown to be more sensitive due to their ability to model nonlinear, individual characteristics [52] [53] [54].

Novel computational models

Although much had been done in what concerns the development of Reinforcement Learning models during the past decades, the concepts inherent to them were only partially applicable to the scope of this study. In fact, this study represents the first attempt to capture the influences of some psychological and cognitive processes which we thought to be involved when performing the training of the AAT.

To model subjects' behavior when performing this 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 section 1 and according to equation 2:

$$w_t(s_i, a_j) = h_t(s_i, a_j) + p(s_i, a_j) + c(s_i, a_j) \qquad (2)$$

Regarding the habit learning component, according to Thorndike's law of exercise that states that an S-R association is strengthened every time the respective stimulus and response are paired [34] and Neal et al., (2006), which brought to light evidence concerning the role of repetition in habit learning [55], we assumed that it would lead to the learning of new S-R associations between the stimulus s_i and the respective instructed action a_j (equation 3)

$$h_{t+1}(s_i, a_j) = (1 + \alpha) \times h_t(s_i, a_j) + \beta \times \frac{0.5}{1 + T(s_i)} \qquad (3)$$

Where α and β are the multiplicative and additive learning rates, respectively, and were constrained to be equal or greater than zero. 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 [56]. 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 [57]. 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 according to equation 4.

$$p(s_i, a_j) = \begin{cases} -\pi \cdot v(s_i), \ if \ a_j = avoid \\ \pi \cdot v(s_i), \ if \ a_j = approach \end{cases} \qquad (4)$$

Where $v(s_i)$ is driven by the rating the subject attributed to that stimulus before initiating the training period. Therefore, it is clear that congruent reactions 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 and it was modeled to represent the degree of cognitive control engaged by subjects (equation 5).

$$c(s_i, a_j) = \begin{cases} C, \ if \ [(v(s_i) < 0) \wedge (a_j = approach) \wedge (instruction = approach)] \vee \\ \qquad [(v(s_i) > 0) \wedge (a_j = avoid) \wedge (instruction = avoid)] \\ 0, \ otherwise \end{cases} \qquad (5)$$

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

Then to transduce these psychological processes into a behavioral measure we resorted to Piéron's Law (equation 6).

$$RT = \alpha I^{-\beta} + \gamma \qquad (6)$$

According to van Maanen *et al.* (2012) Piéron's Law could be used when the discriminability of two competing choices was manipulated (for further details see [58] [59]). Therefore we transformed the preferences into predicted RTs through equation 7.

$$r_t = [w_t(s_i, a_{instructed}) - w_t(s_i, a_{non-instructed}) + k]^{-D} + E \qquad (7)$$

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.

Regarding the enunciated model some notes are important to mention. First, initial conditions for the parameters had to be defined and some constraints had to be imposed on them, otherwise the optimization routine would run into numerical problems all the time. Thus the main constraint which was imposed was that the cognitive control component should be higher than the Pavlovian component (equation 8):

$$C > |\pi \cdot v(s_i)| \qquad (8)$$

With this, we guaranteed that there were not any negative preferences, which, if allowed to exist, would lead to the prediction of imaginary RTs.

Additionally, we thought 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, might subsist even after pre-processing of the data.

A priori hypotheses testing

First, considering prior studies [47] [48] [49], we predicted that the RTs would not be normally distributed. Secondly, besides expecting to partially capture some psychological and cognitive processes of interest, we expected the necessity of having both learning rates since they explain the learning processes that occur in different moments during the task. Thirdly, we predicted the Pavlovian bias to be significantly different than zero and if that was not the case, by using the moving average method we hypothesized that we could reach significance on this parameter.

Thus, to perform a complete study of these hypotheses several computational models were developed (cf. table 1), 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[1]
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 1: Number assigned to the different designed models where the maximum likelihood estimation was performed using different likelihood

[1] This likelihood function was chosen according to the results presented throughout the section 3

functions centered at the predicted RTs [Normal: $RT_i \sim Normal(\widehat{RT_\imath}, \sigma^2)$; Log-Normal: $RT_i \sim LogNormal(Log(\widehat{RT_\imath}), \sigma^2)$].

Initially the models and the processes of optimization and parameter estimation were performed in MATLAB using the *fmincon* function. Notwithstanding, given the fact that these processes were very time-consuming, we decided also to implement the model and respective routines in RSTAN. 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.

Parameter estimation

At the beginning, we assumed that the observed RTs followed a normal distribution centered at the predicted RTs and tried to estimate the parameters using the Ordinary Least Squares (OLS) method through minimization of the residual sum of squares (equation 9).

$$SSE = \sum_{i=1}^{n} (y_i - \widehat{y_\imath})^2 \tag{9}$$

However, despite very common [46] [60] [61], 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 [62].

Therefore, we had to resort to maximum likelihood estimation that allowed the estimation of the optimal parameters of each subject ($\hat{\theta}_s$, cf. equation 10).

$$\hat{\theta}_s = \arg\max_{\theta_s} P(D_s|\theta_s, M) \tag{10}$$

Notwithstanding, we performed the maximum likelihood estimation using two different likelihood functions [Normal: $RT_i \sim Normal(\widehat{RT_\imath}, \sigma^2)$ and Log-Normal: $RT_i \sim LogNormal(Log(\widehat{RT_\imath}), \sigma^2)$].

Consequently, 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.

In an initial phase, we used arbitrary variances for both likelihood functions and then, making use of the predicted data, we estimated the variance which maximized the likelihood for both distributions in order to then accurately compare the results obtained through different computational models.

In order to estimate σ^2 of each observation (RT_i), we first have to convert equation 7 into a regression model, obtaining equation 11:

$$RT_i = (\Delta w_i)^{-D} + E + \varepsilon_i \tag{11}$$

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 ($\widehat{RT_\imath}$) which depend on Δw_i. Therefore, the residual sum of squares (SSE) is given by equation 12:

$$SSE = \sum_{i=1}^{n} (RT_i - \widehat{RT_\imath})^2 \tag{12}$$

Consequently, the residual mean square (MSE) is obtained by equation 13:

$$\widehat{\sigma^2} = MSE = \frac{\sum_{i=1}^{n} (RT_i - \widehat{RT_\imath})^2}{n-p} \tag{13}$$

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

However, this previous reasoning only applies for normally distributed errors [63]. In our specific case, we assumed that the RT_i could also follow a Log-Normal distribution. Therefore, to correctly use the maximum likelihood method we had to find an estimator for the variance of the predicted RTs.

Considering equations 14 (the estimated variance of a normally distributed sample), 15 and 16 (the estimated variance and mean of a log-normally distributed sample, respectively [64]).

$$\widehat{s^2} = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n} \tag{14}$$

$$\widehat{s^2} = \frac{\sum_{i=1}^{n} (log(Y_i) - \hat{\mu})^2}{n} \tag{15}$$

$$\hat{\mu} = \frac{\sum_{i=1}^{n} log(Y_i)}{n} \tag{16}$$

From the analogy found between equation 14 and 15 and the rationale explained to obtain equation 13, we derived our estimator for the σ^2 (equation 17).

$$\widehat{\sigma^2} = MSE = \frac{\sum_{i=1}^{n} (log(RT_i) - log(\widehat{RT_\imath}))^2}{n-p} \tag{17}$$

Model comparison

With the created models, we aimed to validate the quality of the implemented learning approaches (through the use of different model types, cf. table 1), but also the relevance of the Pavlovian parameter. This validation was performed through the application of model comparison procedures to the dataset.

To perform more reliable model comparison, the model evidence (ME) ($P(D|M)$, cf. equation 18) should be used [65] [66].

However, usually, equation 18 is analytically intractable and numerically difficult to compute for the models which we are interested in comparing. Besides usually there is no valid information regarding the prior probabilities of the parameters

$$P(D_s|M) = \int P(D_s|\theta_s, M) \cdot P(\theta_s|M)\, d\theta_s \tag{18}$$

Having this in mind we tried to overcome these difficulties by using approximations to model evidence. These are presented next: the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC)

Both the AIC and the BIC present two main quantities, an accuracy-related term, which is given by the maximum likelihood estimate, and one related to the complexity of the model (cf. equations 19 and 20, where k and n identify the number of parameters of the model and the number of data points to be fitted, respectively).

$$AIC = -2 \cdot (Log(P(D_s|\theta_s, M)) - k) \tag{19}$$

$$BIC = -2 \cdot (Log(P(D_s|\theta_s, M)) - \frac{k}{2} \cdot Log\, n) \tag{20}$$

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 [66].

Therefore, 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.

The comparisons between different models of table 1 were performed through the use of a hierarchical model, which treated the models of our study as random effects (cf. figure 4) [66] [67].

This 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 the unobserved occurrences of the models in the population.

The hierarchical model only needs the model evidences ($P(D_s|M_k)$), which were computed using the approximation of the BIC, and its inversion allows to compute the values of $\boldsymbol{\alpha}$ which, when subtracted the prior, reflect the effective number of subjects for whom a given model generated the data [67].

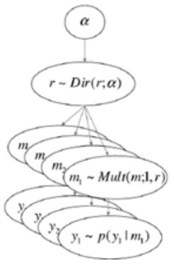


Figure 4: Hierarchical Bayesian model with random effects used to perform Bayesian model selection. Figure adapted from [67].

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 21, where K is the total number of models).

$$\Phi_k = P(r_k > r_{j\neq k}|y), \forall j,k \in K \quad (21)$$

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 [67]. 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 [67]. This quantity was then used to obtain the protected exceedance probabilities (PEPs), through a Bayesian model average of the EPs (equation 22):

$$\widetilde{\Phi_k} = \Phi_k \cdot (1 - BOR) + \frac{1}{K}BOR \quad (22)$$

Thus, these quantities provided the essential information to properly perform the Bayesian model selection (BMS). An important remark regarding equation 22 is that if BOR tends to zero, it means that the hierarchical model of figure 4 is reliably better than chance. Another consequence is, that if that is the case, the PEPs will be very similar to the EPs.

These BMS processes were performed in MATLAB, using the toolbox described in [68], 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 [69].

All results were mainly interpreted in terms of PEPs, as supported in the most recent articles regarding this methodology [66].

3. Results

To study participant's behavioral data obtained from the different versions, we used different methods accordingly to the analysis performed.

Regarding the exploratory data analysis we found evidence to infer that the RTs would be more prompted to be log-normally[2] distributed than normally distributed. This gave an important hint and alerted us not to use the OLS, nor the normal probability density function during the optimization of the computational models' parameters via maximum likelihood estimation. In the following analysis, several results will support this hypothesis.

Model-free analysis

To test whether there was any significant bias *a priori* at a group level the, the RTs acquired from the arrow version of the AAT were analyzed by the following mixed model (equation 23):

$$rt = action * group + (1|subject). \quad (23)$$

The analysis of the mixed model designed above showed that there was no evidence for a main effect of action nor a main effect of group ($p - value > 0.2$). The interaction term showed also evidence of non-significance [$F_{(1,661.11)} = 3.502, p - value = 0.062$]. These results indicated that there was no existent task-related bias.

Therefore, the analysis of the RTs obtained through the assessment version were not corrected. In order to analyze this data we design the following mixed model (equation 24):

$$bias = category * session * trained \\ + generalized + (1|subject) \quad (24)$$

Regarding the variables used, $bias$ represents the difference between the RTs assigned to the avoid action and the RTs assigned to the approach action; $category$ 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 trails with pictures that were not used for generalization, 1 for the trials with pictures used for generalization.

The analysis of this model did not show evidence for any significant main effect, neither for interactions ($p - value \geq 0.150$). These results were not surprising since the bias should not be predicted solely by main effects or 2 way interactions. However the 3 way interaction, which contains information regarding whether or not from one session to another, in a specific condition, there were differences between trained and untrained pictures, was also not significant. 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, we continued with the post-hoc analysis despite the non-significant 3 way interaction. After performing these contrasts we verified that the only significant result was regarding an avoidance bias participants presented towards pictures of the approach neutral condition before training [group

[2] 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.

that did not train (negative group): $z - value = -2.960,\ p - value = 0.003$].

Notwithstanding the above results, we decided to depict 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). (cf. figure 5).

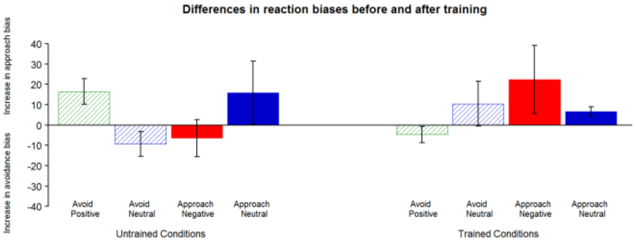


Figure 5: Results from the assessment version of the AAT. The differences in reaction biases before and after training of the positive group, towards negative and neutral pictures they did not train, and for the negative group, towards the positive and neutral pictures they did not train (left side), 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).

According to our hypotheses, we predicted that training one condition would lead participants to show an increased bias towards that condition. From figure 5 we inferred that the majority of the tendencies seemed to go towards the direction of what people trained. Moreover, we continued with the contrast tests and verified that there is a trend for the tendency between the red bars [$z - value = 1.720,\ p - value = 0.090$] and, although not significant, there is a tendency between the green bars pointing to the desired direction [$z - value = -1.580,\ p - value = 0.120$]. Besides, since we wanted to test for generalization effects in behalf of the condition trained, *i.e.*, to verify if participants generalize the learning they had for similar pictures we decided to perform pairwise comparisons between pictures trained and pictures used for generalization. In fact, a trend was obtained for the negative condition [$z - value = 1.890,\ p - value = 0.060$], while the positive only presented a visual tendency towards the hypothesized direction (increase in avoidance bias) [$z - value = -1.540,\ p - value = 0.120$].

Figure 6 allows to visually inspect these results.

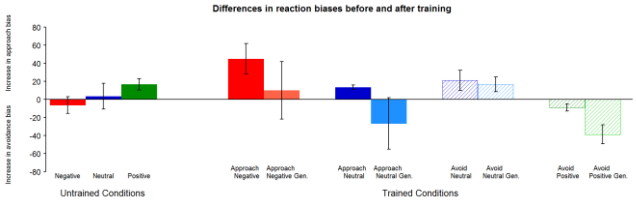


Figure 6: Results from the assessment version of the AAT. The differences in reaction biases before and after training of the positive group, towards negative and neutral pictures they did not train, and for the negative group, towards the positive and neutral pictures they did not train (the three initial 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).

Regarding the training version of the AAT, we decided to concatenate the data-sets of different days for each subject. This procedure was supported by findings provided by [15] [39] [70] and by participants reports stating they had remembered what they had trained, evidencing that besides habit learning there was also a strong evidence for episodic memory. Then, to understand the dynamics of the RTs and the influences of the subjects, we used the mixed models approach.

The predictors used were: $trial$, a continuous variable that represents the trial number, which maximum value is 600; $cond$ is a categorical variable with four levels: 1 for *approach negative* condition, 2 for *avoid neutral* condition, 3 for *avoid positive* condition and 4 for *approach neutral* condition.

Thus, several models, using three trends (linear, exponential and power law), 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. Of all the models created, we verified that the linear fits were the poorest and that the approximation applied on the RTs (logarithm transformation) proved to be a good one, due to the lower LLH values obtained when compared to the linear fits. Considering the models fitted by the exponential and power law fits, we also concluded that the better models required all the random effects, since they had better and congruent AIC, BIC and LLH values.

Nonetheless, when we used the AIC and the BIC to select the best model using all the random effects they were no longer in agreement. Therefore, we decided to analyze the most complex model for both fits because we were interested in verifying the significance of all the predictors and since both were reported in prior studies and so there was no clear preference for one of them. The mixed models analyzed were the following (equations 25 and 26):

$$\mathrm{Log}(rt) = \mathrm{Log}(trial) * cond + (1 + \mathrm{Log}(trial) * cond|subject) \quad (25)$$

$$\mathrm{Log}(rt) = trial * cond + (1 + trial * cond|subject) \quad (26)$$

The analysis showed that there are significant main effects of $trial$ and $cond$ for both fits ($p - value < 0.02$). Moreover, the interaction term also presented high significance for both fits ($p - value < 0.001$).

Figures 7 and 8 depict respectively the exponential and power law models' fits to the behavioral data at group level.

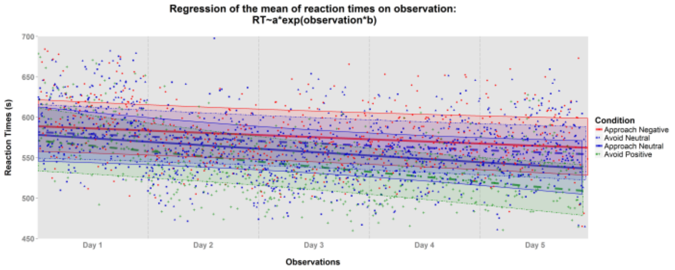


Figure 7: Results of the exponential model's fit performed to the behavioral data at group level.

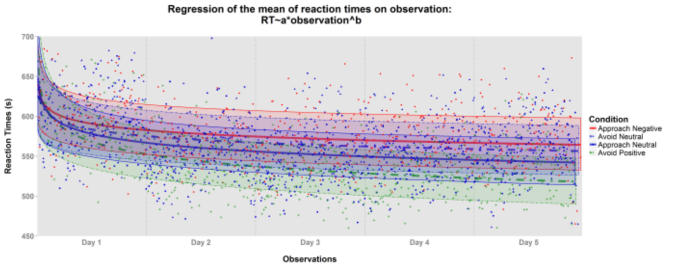


Figure 8: Results of the power law model's fit performed to the behavioral data at group level.

From figures 7 and 8 we verified 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. 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 [$z - value = -2.200,\ p - value = 0.039$].

Besides, we also verified that the starting values of the neutral conditions were not significantly different [$z - value = 0.543$, $p - value = 0.593$], which was interesting because this was according to what was expected. Following this rationale we also decided to test the remaining differences between intercepts and the differences between the slopes. The results showed only a trend for the difference in starting points between the negative and positive condition [$z - value = 1.697$, $p - value = 0.093$], while for the difference in the slopes all comparison showed significance ($p - value < 0.02$), except for the conditions trained by the negative group ($p - value = 0.86$).

These findings indicated that the positive condition was easier to learn when compared to the other conditions and the negative condition appeared to be the most difficult condition to be learned.

Finally we decided to analyze the ratings participants provided during the week of training, since we expected to observe, in the negative group, 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. Moreover, besides the expectancy of participants rating the pictures initially according to their categories before the training, we expected participants to generalize the effects of the training. In order to test these hypotheses two mixed models were design (equations 27 and 28), whose variables' interpretation is similar to the ones of the mixed model design of the assessment.

$$classification = category * session + (1|subject) \tag{27}$$

$$\begin{aligned} classification = category * session * trained \\ + generalized + (1|subject) \end{aligned} \tag{28}$$

The analysis of the first mixed model showed a highly significant main effect of category [$F_{(3,207.91)} = 306.78, p - value < 0.001$] and a significant trend for the main effect of session [$F_{(1,971.99)} = 3.42, p - value = 0.065$], while the term of interaction was not significant ($p - value > 0.20$).

Figure 9 depicts the evolution of the ratings, along sessions, of the pictures trained in the two groups.

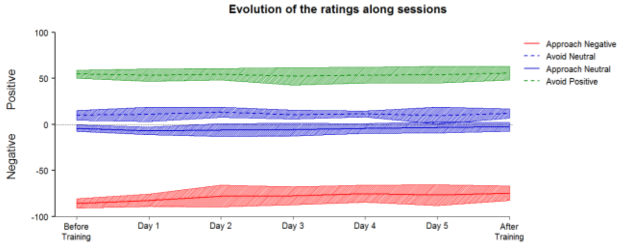


Figure 9: Evolution of the ratings along 5 days of training of the pictures trained in the 4 different conditions.

Although the results above were not clear regarding the possible changes in the ratings, from figure 9 we could notice a slight increase in the average of the ratings of the negative pictures. Therefore, 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 ($p - value > 0.20$). 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.

Concerning the analysis of equation 28 it presented 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 [$F_{(1,1116)} = 1.4, p - value = 0.240$]. The terms of interaction did not presented significance ($p - value > 0.20$). These results were in line with the ones provided by the analysis of equation 27.

Figure 10 depicts the ratings before and after the training for the untrained conditions (left side) and ratings before and after the training for the trained conditions including the generalization (right side).

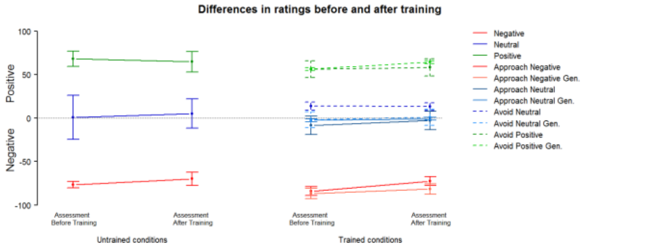


Figure 10: Ratings provided by the participants before and after the training of the positive group, towards negative and neutral pictures they did not train, and for the negative group, towards the positive and neutral pictures they did not train (left side), and ratings before and after the training 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 results of the analysis of equation 28 and visual conclusions from figure 10, post-hoc contrasts were computed. Besides testing whether the ratings before training yield significance on both sides of the figure 10, we also performed pair-wise comparisons between scores before and after the training, regarding the ratings of the trained and untrained conditions. Regarding the first tests the results showed high significance, *i.e.*, the ratings of negative and positive pictures were clearly different from zero, while the ratings of neutral pictures was not (positive: $z - value = 13.87$, $p - value < 0.001$); negative: $z - value = 20.28$, $p - value < 0.001$; neutral: $z - value = -0.248$, $p - value = 0.800$). Relatively to the pair-wise comparisons 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.

Model-based analysis

BMS between the eight models described in table XX yielded a PEP over 95% for the fourth model.

Figure 11 depicts the outcome of the BMC between the models described in table XX.

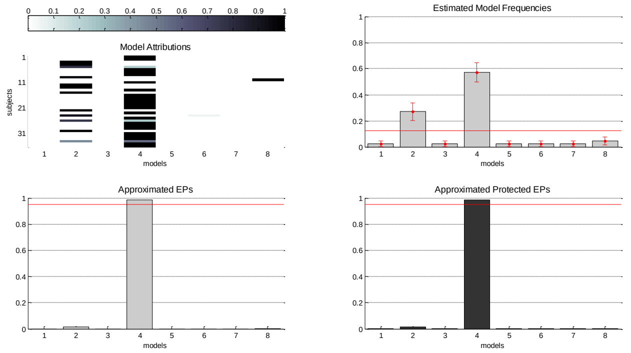


Figure 11: Results of BMC between the models described in table 1 obtained using the BIC, for the population. (Top Left) Model Attribution at subject level, the first eighteen subjects belong to negative group while the last eighteen to the positive group. (Top Right) Estimated model frequencies. (Left and Right Bottom) Approximated Exceedance Probabilities (EPs) and Protected EPs of the eight models.

From figure 11 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.

Then, we proceed with the analysis of the Pavlovian parameter, which was performed considering the parameters estimated for all subjects through the model assigned with the higher value of PEPs. After testing the sample's parameter for normality using the *Kolmogorov-Smirnov* test ($p-value < 0.001$), 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 [65]. Even though the result was not significant, it presented a trend ($p-value = 0.066$).

Nonetheless, a detailed analysis of figure 11 led us to believe that we could have outlier candidate subjects. Therefore a histogram of the parameter's sample was performed (figure 12)

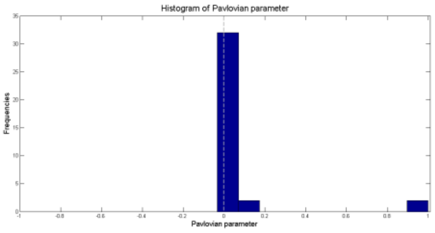


Figure 12: Histogram of the Pavlovian parameter estimated by the computational model selected through BMS. The vertical dashed grey line identifies the value zero.

From figure 12 we verified that the 4th and the 8th subjects of the positive group (subjects 22 and 26 respectively) were two outlier candidates, since they were assigned with a Pavlovian parameter close to 1.

In fact, due to the constraint imposed on the cognitive control component (cf. equation 8) we expected a high multicollinearity between these two components, which from the analysis of the equation that rules the preferences (cf. equation 2) 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 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. After verifying the individual ratings for these two participants, we observed that the subject 22 was in the condition above mentioned. Moreover this subject's behavior was better described by the simplest model. 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. So, 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 ($p-value = 0.160$).

Thereafter, 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).

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 13).

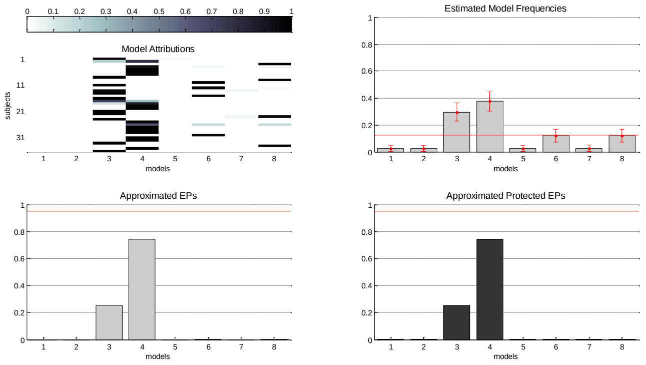


Figure 13: Results of BMC between the models which received as input transformed RTs (via moving average filtering) described in table 1 obtained using the BIC, for the population. (Top Left) Model Attribution at subject level, the first eighteen subjects belong to negative group while the last eighteen 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 13 we observed that the number of subjects whose parameters were estimated by the third model increased a lot relatively to figure 11. 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.

Then, using the above described procedure, we started by testing the normality of the sample's parameter through the *Kolmogorov-Smirnov* test ($p-value < 0.001$) and next 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 ($p-value = 0.036$).

4. Discussion and conclusions

Considering these findings we concluded that this study provided evidence that internal conflicting mechanisms existed 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.

This conclusion was supported by several results. For instance, in the assessment version we were able of partially measuring the unintentional valence processing of each subject, before after the train they went through during 5 consecutive days: 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.

Another finding that supports this was 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.

However, due to factors such as the great inter-subjects variability and to design's issues inherent to this version of the task, we only obtained significant trends and tendencies pointing to the desired direction.

Regarding the training version of the AAT we verified that participants of both groups learned the trained conditions, showing evidence that repetition actually plays a key role in habits formation [38] [39] [40]. Besides this, we also found evidence for automaticity of the trained behaviors due to the fact that there were conditions that reached an asymptotical reaction time. Moreover, we also found evidence for significantly stimuli's valence processing modification, specifically for 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.

Then to study the psychological and cognitive processes of the training behavioral data, novel computational learning frameworks were compared and selected through BMC. In fact, 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. This process was performed by using the BIC as model evidence's approximation.

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 [57], 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 was 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 a 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, it was the fact that we found consistent and coherent findings in both of the performed analyses.

Nonetheless, the accurate use of hierarchical models of parameter estimation would be very beneficial since these models explicitly deal with the within-subject variability on the parameter estimates.

Regarding model comparison, the use of better model evidence's approximations (computed through Markov Chain Monte Carlo sampling [71] or variational Bayes techniques [66] [67]) should also be targeted, in order to formally validate model selection procedures.

Acknowledgements

I would like to thank all my friends and family for all the support and everything else. It has truly been a pleasure to study Biomedical Engineering at IST.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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).

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

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

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

[25] 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).

[26] 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).

[27] 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).

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

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

[30] 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.

[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).

[32] Domjan, M., Principles of Learning and Behavior Active Learning. 6 ed. 2010: Cengage Learning.

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

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

[35] 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.

[36] 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.

[37] 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.

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

[39] 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.

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

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

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

[43] 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.

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

[45] 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.

[46] 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.

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

[48] 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.

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

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

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

[52] 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.

[53] 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.

[54] 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.

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

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

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

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

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

[65] 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.

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

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

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

[69] 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).

[70] 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.

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