Insights into reward-based decisions
using computational models and ultra-high field MRI.

Insights into reward-based decisions using computational models and ultra-high field MRI.

Inauguraldissertation

zur

Erlangung der Würde

einer Doktorin der Philosophie

vorgelegt der

Fakultät für Psychologie

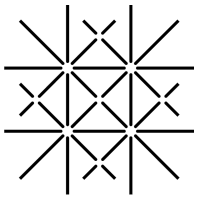
der Universität Basel

von

Laura Fontanesi

aus Peschiera del Garda (VR), Italien

Basel, 2018



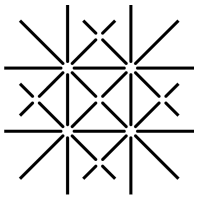
Genehmigt von der Fakultät für Psychologie
auf Antrag von

Prof. Dr. Sebastian Gluth
Dr. Stefano Palminteri

Basel, den _____

_____

Prof. Dr. Alexander Grob



Declaration

I, Laura Fontanesi (born August 8th, 1989 in Peschiera del Garda VR, Italy) hereby declare the following:

(1) My cumulative dissertation is based on three manuscripts. I contributed substantially and independently to all manuscripts in this dissertation. In particular, I was primarily responsible for all data analyses and for the writing of all manuscripts. In addition, I was primarily responsible for data collection in the second and third manuscripts. I was jointly responsible for the ideas in all manuscripts.

(2) I only used the resources indicated.

(3) I marked all the citations.

Basel, November 30th, 2018

Laura Fontanesi

Acknowledgements

I would like to thank my supervisors, Jörg Rieskamp and Sebastian Gluth, for guiding me through my PhD, for teaching me how to be clearer and more confident when presenting my work, and for always providing very instructive feedback. I am also grateful to Birte Forstmann, who supervised me during my Master Thesis and welcomed me back during my PhD Mobility Fellowship in Amsterdam. Thanks to Maël Lebreton and Stefano Palminteri, for trusting me with their cool dataset and accepting my obsession for Bayesian statistics. Thanks to my collaborator and fellow PhD student Mikhail Spektor, for indulging such obsession and always setting high quality standards for methodological practices as well as for meeting deadlines on time. It was a pleasure working with all of you and I hope this work will extend to future collaborations.

During this roller-coaster journey that is a PhD, I was lucky to be surrounded by funny and inspiring PhD students and Post-Docs, that contributed to the highs of it: Mikhail, Janine, Sebastian, Ollie, Marin, Nathaniel, Jana, Ash, Tehilla, Dimitris, Rebecca, Regina, Peter, and Anne. Thanks for being such cool colleagues! And thanks to the other two members of the Gelman's book club: Sebastian and Mikhail. I will always carry those instructive memories in my heart.

I was double-lucky to be supported by my wonderful family and by my partner, Gilles. Thanks for always be there for me, for celebrating the good and helping me seeing the positive side of things – which, I admit, I sometimes fail to do. Thanks to my dad, for always showing interest in my work and teaching me to never forget about the "bigger picture". And, Gilles, thanks for your tireless support, and for reminding me every day of the things that are the most important in my life.

Finally, "grazie di cuore" to Jana, Ash, and Gilles, for helping me revising the framework of this dissertation. You will be rewarded with Italian cuisine that will hopefully meet your (high?) expectations.

Abstract

The ability to integrate past and current feedback associated with different environmental stimuli is crucial for adaptive and goal-directed behavior. The field of reinforcement learning (RL) focuses on understanding such ability: Computational models propose algorithms for the updating of beliefs based on the received feedback, while neural models describe how these algorithms are implemented in the brain. In this dissertation, I investigate learning-by-feedback both at the algorithmic level (in the first part) and at the implementation level (in the second part).

In the first part, in particular, I focus on human behavior during learning-by-feedback in the presence of both monetary gains and losses (first manuscript) or of different magnitudes of monetary gains (second manuscript). To date, most computational RL models focused on trial-by-trial dynamics and have not explained how response times (RTs) change during learning or are affected by different learning contexts (e.g., in the presence of gains vs. losses). In both manuscripts, I argue that RTs are crucial for the understanding of reward-based decisions: In both studies, participants' RTs were affected by different learning contexts, while their choice preferences not always were. To jointly explain the effects on both preferences and RTs, I used a sequential sampling model, the diffusion decision model (DDM). In the first manuscript, I propose a meta-analytical approach to simultaneously analyze the effects of different learning contexts on choice preferences and RTs from four independent experiments. In the second manuscript, I propose a new model that incorporates an RL algorithm into the DDM.

In the second part of my thesis, I investigated the coding of losses and gains by the dopaminergic nuclei in the human brain. Since these nuclei are situated deep in the brain, their signal is hard to study using non-invasive imaging techniques such as magnetic resonance imaging (MRI): To date, human studies have provided incomplete and partially contradicting findings about the reward signals in dopaminergic nuclei. In the third manuscript, I provide evidence that clarifies the role of the dopaminergic nuclei when receiving more or less surprising gains and losses, as well as when expecting higher or lower outcome risk. To do so, I capitalize on ultra-high field MRI and on the use of multimodal images to delineate the dopaminergic nuclei on a participant level.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

> The man who has fed the chicken every day throughout its life at last wrings its neck instead, showing that more refined views as to the uniformity of nature would have been useful to the chicken.
>
> *Bertrand Russell*
> *The Problems of Philosophy*

As human beings, we constantly process streams of inputs coming from our sensory organs: molecules in the air and food become smells and tastes, waves of particles become images and sounds. All these inputs are decoded as sensory information that is transmitted to our brain and helps us to avoid obstacles and dangers, as well as to get to resources that keep us satisfied or make us thrive. However, information comes with some level of uncertainty, in this case, sensory uncertainty. When picking wild mushrooms, it is very important to correctly discriminate between poisonous and non-poisonous ones based on their appearances, though it might not be always easy because of natural variability. On top of this, the environment itself is highly dynamic. Even after being certain of the identity of a tree, its value (e.g., the number of apples it delivers) may change in time, because of seasonal changes or calamities of a different nature. Through the past decades, cognitive psychologists and neuroscientists have been interested in understanding how humans are not only able to process uncertainty at these different levels (i.e., perception and evaluation), but also to integrate past and current information and use this knowledge to

guide their behavior. While the first type of uncertainty – the perceptual one – has been the focus of psychophysics and *perceptual decision making*, the second type – the one about value – has been the focus of behavioral economics and *value-based* (or economic) *decision making*. Recently, more and more researchers have tried to understand the common mechanisms underlying both perceptual and value-based decision making, at a behavioral and at a neural level (Dutilh & Rieskamp, 2016; Polania, Krajbich, Grueschow, & Ruff, 2014; Turner, Schley, Muller, & Tsetsos, 2018). Computational models of decision making have been an important methodological tool that has helped bridging this gap, as they decompose behavioral measures into latent psychological constructs that can be compared across domains and mapped to neural activity (Summerfield & Tsetsos, 2012).

In this cumulative dissertation, I focus on reward-based decisions, a particular kind of value-based decision. Rewards – as well as punishments – are not defined by their physical properties but by what they induce (Schultz, 2015): either approaching (for rewards) or avoiding (for punishments) behavior. Depending on whether someone enjoys the taste of apples coming from a certain tree, they will be more or less likely to go back and pick another apple from that tree in the future. Moreover, by sampling a few apples from each tree in a garden, they can learn which tree consistently gives better apples and which tree is to be avoided. Throughout experience, one can thus learn to maximize rewards and minimize punishments. Research in neuroscience tells us that the human brain – as well as the brain of other animals – is wired for this: While certain areas in the brain encode value and deviations between previous expectations and current experiences, other areas integrate these signals to guide and regulate actions (Haber & Knutson, 2010).

The first part of this dissertation focuses on the cognitive mechanisms underlying reward-based decisions. In order to better understand such mechanisms, I used computational models from the perceptual decision making tradition, i.e., sequential sampling models (SSMs). SSMs describe within-trial dynamics, i.e., how the decision evolves from stimuli presentation to the commitment to an answer. As a consequence, they make predictions on both choice preferences and response times (RTs). In this framework, the same decision problem (e.g., which one of two visual stimuli is brighter?) is presented several times to participants, and each choice is treated as an independent observation of the same noisy decision process. These models are fundamentally different from traditional models of reward-based decision making, such as reinforcement learning (RL) models. RL models describe trial-by-trial dynamics, i.e., how choice preferences change as a function of the experienced feedback. Because they make no assumptions on within-trial-dynamics, they only

predict choices and not RTs. In the first manuscript (Chapter 2), I used a sequential sampling model (SSM) – the diffusion decision model (DDM Ratcliff, 1978; Ratcliff & Rouder, 1998) – to understand the mechanisms underlying learning to choose the most advantageous of two winning options (yielding monetary gains) or two losing options (yielding monetary losses), in the presence of more (full feedback) or less (partial feedback) information. In the second manuscript (Chapter 3), I proposed a new computational model that integrates the RL algorithms – describing the updating of beliefs after receiving feedback – to the DDM. While a similar approach has been proposed before (Pedersen, Frank, & Biele, 2017), I show how our model can better explain behavior in different learning contexts: across different levels of difficulty and across contexts yielding higher or lower monetary gains. I will further explain these two manuscripts and their unique contribution to the literature – both theoretical and methodological – in Section 1.1.

The second part of the dissertation focuses on the neural mechanisms underlying reward-based decisions. In particular, I investigated the areas in the brain that encode the reward signal and transmit it to important areas of decision making. These areas are situated in a evolutionary older part of the brain – the midbrain – and transmit the reward signal by means of the dopamine neurotransmitter. Most of the knowledge we have concerning these areas comes from animal studies. This is because these areas are situated far from the skull and their signal is difficult to measure without highly invasive techniques, such as microelectrode recordings. In the third manuscript of this dissertation (Chapter 4), we measured signal from dopaminergic nuclei in the human brain. In order to more reliably estimate the signal from these areas, we used ultra high field magnetic resonance imaging (UHF-MRI) combined with multimodal imaging to delineate the dopamine nuclei at an individual level. In this study, participants engaged in a gambling task in which they could experience monetary gains and losses and different degrees of uncertainty about the outcomes. In Section 1.2, I will explain how these results help clarifying some ambiguous findings in previous human studies, as well as extending previous findings in non-human animals samples to a human sample.

1.1 Computational models of decision under uncertainty

Cognitive psychology is the study of the cognitive processes that underlie human and animal behavior. In controlled settings (e.g., in a laboratory study), participants are

conditioned to different treatments and their behavior is measured as a function of them. Significant changes in behavior due to experimental manipulations are therefore interpreted as a consequence of latent psychological, cognitive constructs. If we are interested in, for example, the effect of reward probability on behavior, we can present participants with choices between options associated with low gain but high probability and options associated with high gain but low probability. By observing participants' choice preferences, we interpret their preference towards the high probability, lower gain option as a preference for certain outcomes, or risk-aversion. However, this method does not allow us to quantify a participant's risk preference and compare it to others'. This method is also limited as it does not allow us to make predictions for new stimuli. For instance: What happens if we double both options' reward? What happens if we lower both probabilities? To be able to do so, what is missing is a mathematical function that links magnitude of outcomes and their probability of occurring to the probability of choice of one option over another. Ideally, this function's parameters can also (1) be tuned, or estimated, based on observed performance and (2) be related to the latent psychological, cognitive constructs that we want to study. In this way, it is possible to observe a participant's performance, estimate and quantify their latent cognitive processes, and make predictions for behavior when new, unseen, stimuli are presented. Computational models serve exactly these purposes, by describing cognitive or neural processes by means of mathematical functions and defining probability distributions over the parameters based on observed data (Lewandowsky & Simon, 2010).

Providing an overview of computational models of decision under uncertainty is beyond the scope of this dissertation (for this purpose see, e.g., Glimcher & Fehr, 2014; Busemeyer, Wang, Townsend, & Eidels, 2015). In what follows, I will focus on two particular classes of models: SSMs, that are prominent models in perceptual decision making, and RL models, that are prominent models for reward based decisions. Finally, I will discuss work aimed at bridging the gap between these two classes of models.

1.1.1 Sequential Sampling: the diffusion decision model

Within the sequential sampling framework, making decisions between two options is similar to Bayesian hypothesis testing (Gold & Shadlen, 2007): the deliberation process starts with some prior (i.e., one option is better or they are the same), some evidence is then accumulated, and the balance in favor of one or the other alternative (i.e., the decision variable) changes accordingly. The process ends when the evidence in favor of one

alternative reaches a certain criterion value (i.e., the decision rule). Since evidence is noisy, the more evidence is accumulated, the more the decision variable is stable, reflecting the true state of the world. Different SSMs can be more or less similar to Bayesian hypothesis testing, depending on specific assumptions they make about how evidence is accumulated (Smith & Ratcliff, 2004; Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006). First of all, in Bayesian hypothesis testing, evidence is accumulated as a single total, and specifically as the ratio of posterior probabilities (i.e., the probability of an hypothesis being true given prior odds and the evidence). While random-walks models are SSMs that maintain this assumption, accumulator or counter models assume that evidence for two or more alternatives is accumulated in separate sums. Examples of random-walks models are the DDM (Ratcliff, 1978; Ratcliff & Rouder, 1998) and the Ornstein-Uhlenbeck diffusion model (O-U, Busemeyer & Townsend, 1993), while examples of accumulator models are the Poisson counter model (LaBerge, 1994) and the leaky competing accumulator model[1] (LCA, Usher & McClelland, 2001). Moreover, in Bayesian hypothesis testing, evidence is accumulated in discrete time steps, as in the Poisson counter model, whereas other SSMs allow evidence to be accumulated continuously in time, such as the DDM, O-U, and the LCA. Finally, in Bayesian hypothesis testing, no evidence is ever discarded. Certain SSMs, such as the LCA, relax this assumption to account for more biologically plausible processes in which older information is discounted, i.e., "leaks".

Despite the many differences across the SSMs that have been proposed in the last decades, what they all have in common is that they describe the deliberation process that leads to a single decision, usually in the order of a few seconds) as bounded accumulation of noisy evidence. This notion has high explanatory power at a behavioral and at a neural (see section 1.2) level.

At a behavioral level, SSMs make probabilistic predictions for both choices and RTs. This feature separates SSMs from other prominent models of noise integration that are limited to choices, such as signal detection theory (Green & Swets, 1966), and allows them to explain correlations between RT and accuracy. One of the most robust findings in decision making research is the speed-accuracy trade-off (Heitz, 2008; Luce, 1986), i.e., the finding that speedy decisions are also less correct, while slow decisions tend to be more correct. This effect can be induced in the lab by asking participants to be either very fast or very accurate in different experimental conditions. With standard statistical analyses (i.e., t-tests or

[1] The LCA can reduce to the O-U model in the presence of not more than two alternatives

ANOVAs), one can measure statistical differences between conditions, separately on choices and RT. SSMs not only predict such behavioral patterns, but (unlike standard statistical tests) they also offer a mechanistic explanation and allow us to quantify latent cognitive variables that are assumed to cause this effect. In the case of the speed-accuracy trade-off, this latent variable is cautiousness, and is formalized in SSMs in terms of the criterion value at which evidence accumulation effectively stops (i.e., the *decision threshold*). At equal incoming rates of evidence, lower threshold values produce faster responses but, since less evidence has been accumulated, the decision is also more affected by noise. Another well established result is the difficulty effect (Ratcliff & Rouder, 1998), i.e., the finding that, when stimuli are more discriminable, responses are faster and more accurate. In this case, speed and accuracy are positively correlated. SSMs offer a mechanistic explanation of this effect: For higher accumulation rates of evidence, assuming that the noise in the signal is constant, the signal-to-noise ratio increases, and evidence in favor of the correct option reaches the correct decision threshold faster.

The DDM is widely used in decision making, and has a relatively high mathematical tractability[2] . According to the DDM (Figure 1.1), evidence is accumulated continuously and constantly over time, without leakage, in a single sum, and the decision threshold remains fixed within trials. Similarly to Browninan motion, the noise in the evidence is normally distributed and temporally uncorrelated. The within-trial accumulation of evidence follows the following equation:

$$x_{i+1} = x_i + \mathcal{N}(v \cdot dt, \sqrt{dt}), x_0 = a/2 \tag{1.1}$$

where x_i is the accumulated evidence at iteration i (i.e., the decision variable), dt is the integration time unit (which approaches 0 in the limit, corresponding to continuous time), v is the mean of the incoming evidence (i.e., the accumulation or *drift* rate), a is the decision threshold, and x_0 is the prior evidence before evidence accumulation begins (i.e., the *starting point*). A response is initiated when x reaches either the upper threshold ($x \geq a$), in which case the response is correct, or when x reaches the lower threshold ($x \leq 0$), in which case the response is incorrect. The RT is given by the number of iterations before the threshold is reached [3] plus a quantity referred to as the *non-decision time*. The non-decision time

[2] Most SSMs (but see S. D. Brown & Heathcote, 2008) do not have a likelihood function, and can only be fitted using simulation-based methods. These methods can significantly increase the time needed for parameter estimation.

[3] To make realistic predictions of RTs, the dt should be a quantity close but not equal to 0, e.g., $dt = 0.0001$ seconds.

corresponds to the time during which evidence is not accumulated, and usually accounts for motor or stimulus encoding processes. In this notation, responses are coded as correct and incorrect. However, depending on the particular paradigm, it is possible to code responses differently, e.g., right and left. In such case, the starting-point can indicate a prior in favour of either right or left options.

Like other SSMs, the DDM accounts for the speed-accuracy trade-off by means of the threshold parameter (with a lower threshold corresponding to speedy and inaccurate choices) and for the difficulty effect by means of the drift-rate parameter (with a lower drift-rate corresponding to difficult decisions) (Ratcliff & Rouder, 1998). The DDM also accounts for other behavioral effects such as post-error slowing (e.g., Dutilh et al., 2012) and bias effects (Ratcliff, 1985; Leite & Ratcliff, 2011; Mulder, Wagenmakers, Ratcliff, Boekel, & Forstmann, 2012). In its more complete form, the DDM includes across-trial variability in the parameters, to account for observed effects on the tails of the RT distributions (Ratcliff & Rouder, 1998), although the parameters that describe across-trial variability cannot always be reliably estimated (Boehm et al., 2018). The DDM – in its simpler and more complex forms (i.e. without and with trial-by-trial variability) – has been used to account for behavioral effects on choices and RTs across psychological domains and in different tasks (for an overview, see Ratcliff, Smith, Brown, & McKoon, 2016).

1.1.2 Reinforcement learning models

The modern field of RL started in the 1980s, when researchers in animal behavior, artificial intelligence, and operations research came together to formalize the problem of learning-by-feedback (Sutton & Barto, 1998; Wiering & vanOtterlo, 2012). Feedback can be of different nature, depending on whether it identifies a correct answer or whether it merely evaluates an action. The first kind of feedback is used in *supervised* learning. This kind of learning is mainly studied in the field of artificial intelligence, where artificial agents are trained to, e.g., classify objects based on their physical features. In RL framework, however, the feedback is always evaluative, therefore learning is *unsupervised*. In RL problems, the goal of an agent is to learn by trial-and-error to choose the options that maximize the feedback. In human studies, to motivate this process, feedback is often translated into monetary rewards. Moreover, in its general form, the RL problem refers to settings in which an agent has to predict the outcome of an action, conditional on a particular state of the environment (i.e., *associative* or *model-based* learning). In this thesis, I will only consider

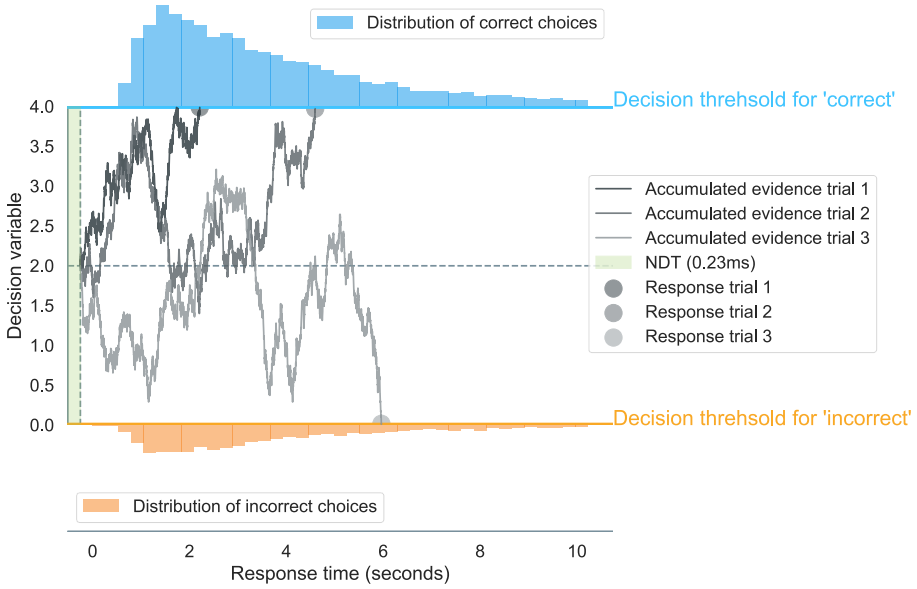


Figure 1.1: The diffusion decision model is a sequential sampling model that describes within-trial dynamics. The trajectories in the middle of the figure represent the accumulated evidence in three different hypothetical trials. In all trials the drift-rate is the same but, because of within-trial noise, the decision variable evolves differently, and the decision outcome depends on which decision threshold is reached first. On the top and bottom of the plot, the distribution of response times (RT) for correct and incorrect trials is shown. Because the drift-rate is positive, the majority of responses are correct. The model also predicts skewed RT distribution, as is often found in human data.

the case in which agents learn to act in one situation and the outcome of the action is not conditional to the state of the environment (i.e., *non-associative* or *model-free* learning). A typical example of such setting is the n-armed bandit problem. In this problem, an agent repeatedly chooses from n different options, and a numerical reward is provided after every choice. These rewards are drawn from distributions that are unique to that option and are not known by the agent. Different RL paradigms can differ in terms of, e.g., the shape of the distribution from which the outcomes are sampled (e.g., normal distribution, binomial distribution), whether the mean of the distribution changes throughout the experiment (i.e., *dynamic* environment) or not (i.e., *stationary* environment), or whether the feedback corresponding to the unchosen options is provided after each choice (i.e., *full* feedback) as opposed to receiving only the feedback associated with the chosen option (i.e., *partial* feedback).

The way this learning process occurs is described by algorithms that take as input the old expectations and a newly experienced outcome, and give as output a new, updated, expectation. These algorithms are often referred to as *learning rules*. A key variable in all learning rules is the reward prediction error (RPE) δ, which measures the difference between the expected and realized rewards (Niv & Schoenbaum, 2008):

$$\delta_t = f_{A,t} - Q_{A,t-1} \tag{1.2}$$

where $f_{A,t}$ is the feedback received from option A at trial t, and $Q_{A,t-1}$ is the expectation of option A in the previous trial. Only when the error is higher or lower than zero (i.e., if the expectations were either too optimistic or too pessimistic) the expectations need to be updated. A simple way to update such expectations is by computing a weighted average of the old expectation and the RPE, according to the Rescorla-Wagner rule (Rescorla & Wagner, 1972):

$$Q_{A,t} = Q_{A,t-1} + \alpha \cdot \delta_t \tag{1.3}$$

where α ($0 \leq \alpha \leq 1$) is the *learning rate* parameter. When α is low, the highest weight is given to prior expectations: learning is slower but more stable (Figure 1.2, left column). On the other hand, when α is high, the highest weight is given to the updated information: learning is faster but more subject to noise in the feedback (Figure 1.2, right column). Learning rates can be fixed throughout learning or change as a function of the trial number (e.g., Yechiam & Busemeyer, 2005), or of the unsigned RPE (e.g., Pearce & Hall, 1980; Diederen & Schultz, 2015). On top of this, previous studies have shown that separate learning rates might be necessary for positive and negative RPEs, to account for asymmetries

in the way negative and positive errors are processed (e.g., Gershman, 2015; Niv, Edlund, Dayan, & O'Doherty, 2012; Frank, Moustafa, Haughey, Curran, & Hutchison, 2007). Other studies (e.g., Palminteri, Khamassi, Joffily, & Coricelli, 2015) proposed separate leaning rates for chosen and unchosen options, since more attention might be given to the feedback corresponding to the chosen option, or for a particular context (Palminteri et al., 2015).

In order to predict the behavior of an agent, a second function has to map the expected rewards to the probability of choosing one over the other options. These functions are often referred to as *decision rules*. A widely used decision rule is the *softmax* rule (Luce, 1959; Bridle, 1990):

$$p_{A,t} = \frac{e^{\theta Q_{A,t}}}{\sum_{j=1}^{n} e^{\theta Q_{j,t}}} \tag{1.4}$$

where $p_{A,t}$ is the probability of choosing option A at trial t, $Q_{A,t}$ is the expected reward of option A at trial t, j are the remaining n options, and θ ($\theta \geq 0$) is the *sensitivity* parameter. When θ is low (i.e., approaching 0), choices become more random (i.e., less sensitive to differences in Q values between the options). Conversely, when θ is high, choices become more deterministic (i.e., more sensitive to Q value differences). More deterministic behavior can be efficient in situations that do not require exploration among the different options (Sutton & Barto, 1998).

RL models have been successful in explaining how choice preferences evolve during learning-by-feedback, and their increased popularity in the last decades is likely linked to the discovery of a RPE signal in the brain (see Section 1.2). In behavioral research, they have been extensively used to differentiate between healthy and clinical populations (e.g., Frank, Seeberger, & O'Reilly, 2004; Waltz, Frank, Robinson, & Gold, 2007; Maia & Frank, 2011) or across age groups (Palminteri et al., 2015; Christakou et al., 2013). An extensive body of literature has seen the application of RL models to the IOWA gambling task, i.e., a learning task that is used as diagnostic tool in the clinical field (Busemeyer, Stout, & Finn, 2003; Yechiam & Busemeyer, 2005; Worthy, Hawthorne, & Otto, 2013). However, the reliability of RL models parameters to discriminate between healthy and clinical population has also been questioned (Steingroever, Wetzels, & Wagenmakers, 2013, 2014). Recently, RL models have been extended to explain context effects (Spektor, Gluth, Fontanesi, & Rieskamp, in press).

Figure 1.2: Reinforcement learning models describe trial-by-trial dynamics. The same feedback (first row) is given to two different agents, after choosing between two options. While the feedback comes with some noise, on average one option (i.e., the correct option) yields higher feedback than the other (i.e., the incorrect option). Throughout experience, the agents update their expectations (i.e., the Q values associated with the two options). The agent with a low learning rate is more conservative: its estimates change more slowly and are more stable in time. Finally, higher feedback expectations towards one option predict higher chance to choose that option: The conservative agent's choices are more consistent across trials.

1.1.3 Combining models to explain within and trial-by-trial dynamics

Traditionally, computational models of decision making have not been shared across the perceptual and the economic domains. In the last decades, however, SSMs such as the DDM have been applied more and more on value-based decisions (e.g., Polania et al., 2014; Polania, Moisa, Opitz, Grueschow, & Ruff, 2015; Clithero, 2018; Milosavljevic, Malmaud, Huth, Koch, & Rangel, 2010; Cavanagh et al., 2011; Cavanagh, Wiecki, Kochar, & Frank, 2014; Frank et al., 2015; Ratcliff & Frank, 2012). Specific SSMs for value-based decision making have also been proposed (e.g., Busemeyer & Townsend, 1993), as well as an extension of the DDM that capitalizes on eye-tracking data to explore the role of attention in value-based decisions (Krajbich, Armel, & Rangel, 2010). The trend of applying SSMs to value-based decisions started from the observation that, despite the different nature of the information that is processed in perceptual and in value-based decision making, the way noisy information is integrated in order to deliver an often binary output (e.g., accept or reject A, choose A over B) is strikingly similar. For example, when choosing under time pressure between options based on either their appearances or value, people tend to be less accurate in their choices (e.g., Ratcliff & Smith, 2004; Heitz, 2008; Milosavljevic et al., 2010). This phenomenon is explained in both domains by changes in the decision threshold. Another example is when choosing between two very similar options (i.e., with similar appearances or value): in this case, people tend to be slower and also less accurate (e.g., Bogacz et al., 2006; Ratcliff et al., 2016; Busemeyer & Townsend, 1993; Cavanagh et al., 2014; Krajbich et al., 2010; Clithero, 2018). This phenomenon is explained in both domains by changes in the drift-rate.

To date, however, most applications of SSMs in the economic domain have been limited to non-learning contexts. This is because SSMs describe the evolution of the decision trial within a trial, and do not make any assumption on trial-by-trial dynamics. Moreover, in order to infer SSM parameters from the participants' choices and RTs, the same decision problem (e.g., choose the brightest of two visual stimuli) is presented multiple times, and each decision is treated as an independent observation. On the other hand, RL models describe the process of updating beliefs as a function of feedback received in subsequent trials. Because RL models make no assumptions about within-trial dynamics, they only predict choice preferences and not RTs. To infer RL parameters, participants are presented with noisy feedback after each decision, and performance during a learning session – starting with the presentation of stimuli with unknown reward distributions – is

considered as a whole: each response depends on the feedback provided in the previous trials (i.e., the reinforcement history). Therefore, both SSMs and RL models have strengths and limitations: While SSMs make predictions on RTs but cannot explain sequential feedback effects, RL make predictions only on choices but can explain trial-by-trial dynamics during learning-by-feedback.

Recently, however, Frank et al. (2015) and Pedersen et al. (2017) proposed ways to extend the DDM to a learning-by-feedback paradigm. In both studies, a learning rule is used to update options' expectations after receiving feedback: While Frank et al. (2015) assumed that the updating was that of an ideal Bayesian observer, Pedersen et al. (2017) assumed that the updating was described by the Rescorla-Wagner rule (see Equation 1.3). The difference between the options' values was then used to define the drift-rate of the DDM in each trial, as was previously done across-trials in non-learning contexts (see, e.g., Milosavljevic et al., 2010). In Frank et al. (2015)'s study, the threshold of the DDM was proportional to the absolute value of the differences between values, to account for effects of conflict on the threshold parameter found in a previous study (Cavanagh et al., 2014). Their model was further improved by neural correlates – from simultaneous electroencephalogram (EEG) and functional MRI (fMRI) – of value representations and of conflict. On the other hand, Pedersen et al. (2017), proposed a model with separate learning rates for positive and negative RPE, and in which the threshold decreases in time. Therefore, these studies represent first instances of RLDDM, i.e., combinations of RL and DDM. A representation of an RLDDM is shown in Figure 1.3: The choice probabilities predicted by the RLDDM in which the drift-rate is proportional to the differences in learned values are very similar to the choice probabilities predicted by the simple RL (Figure 1.2) – given the same feedback information. At the same time, by describing the within-trial dynamics as bounded accumulation processes (Figure 1.1), the RLDDM can, in addition, predict RTs throughout learning.

These studies provided insights on the way SSMs and RL models can be unified to provide a more detailed account of reward-based decisions. However, both approaches have not been tested in learning contexts: (1) with partial feedback, (2) with higher or lower gain expectations, (3) in the loss domain, (4) with normally distributed (as opposed to binomial distributed) reward distributions. Moreover, by assuming a linear relationship between the differences in values and the drift-rate, they make the prediction that, by providing easier decision problems, the drift-rate can increase indefinitely (e.g., if a value difference of 1 corresponds to a drift-rate of 1, then a value difference of 100 corresponds to a drift-rate of

100).

In the first and second manuscripts of this thesis, I analyzed data from RL tasks with different learning contexts compared to Frank et al. (2015)'s and Pedersen et al. (2017)'s studies. By using three classes of models – RL, DDM, and RLDDM – I present results that challenge some of the assumptions made by the RLDDMs proposed so far, and propose a new RLDDM.

1.1.4 First manuscript: Decomposing the effects of context valence and feedback information on speed and accuracy during reinforcement learning: A meta-analytical approach using diffusion decision modeling

In the first manuscript (Chapter 2), I analyzed data from four independent RL experiments in which participants were presented with different learning contexts. The task was a n-armed bandit task. In each block, a total of 8 new options (identified by symbols, see Figure 2.1) were presented in groups of 2 at a time[4] , and participants had to choose between one of them. While in experiments 1, 2, and 3 the maximum time for responding was 3 seconds, in experiment 4 it was 1.5 seconds. The pairs of options could be made of two losing options (yielding a loss of 1 point with either high or low probability) or two winning options (yielding a gain of 1 point with either high or low probability). Moreover, the feedback of pairs of options could be partial (only the feedback of the chosen option is shown) or complete (both feedbacks are shown). The four experiments differed in the number of learning sessions, trials per learning session, and participants (see Table 2.1).

To test the effect of valence (i.e., losses vs. gains) and feedback information (i.e., complete vs. partial feedback) and of their interaction on accuracy and RTs, we fitted two separate ANOVAs, adopting a Bayesian mixed model meta-analysis approach (Singmann, Klauer, & Kellen, 2014). By doing so, we could test (1) main and interaction effects across experiments, (2) whether these effects were similarly strong in each experiment, and (3) whether participants were more accurate or faster in each experiment. Replicating previous reports (Palminteri et al., 2015; Salvador et al., 2017), we found that participants were slower and less accurate in the partial feedback condition – similarly to a difficulty effect – and that they were slower in the loss domain. There was also an interaction effect of feedback

[4] The options pairs were fixed.

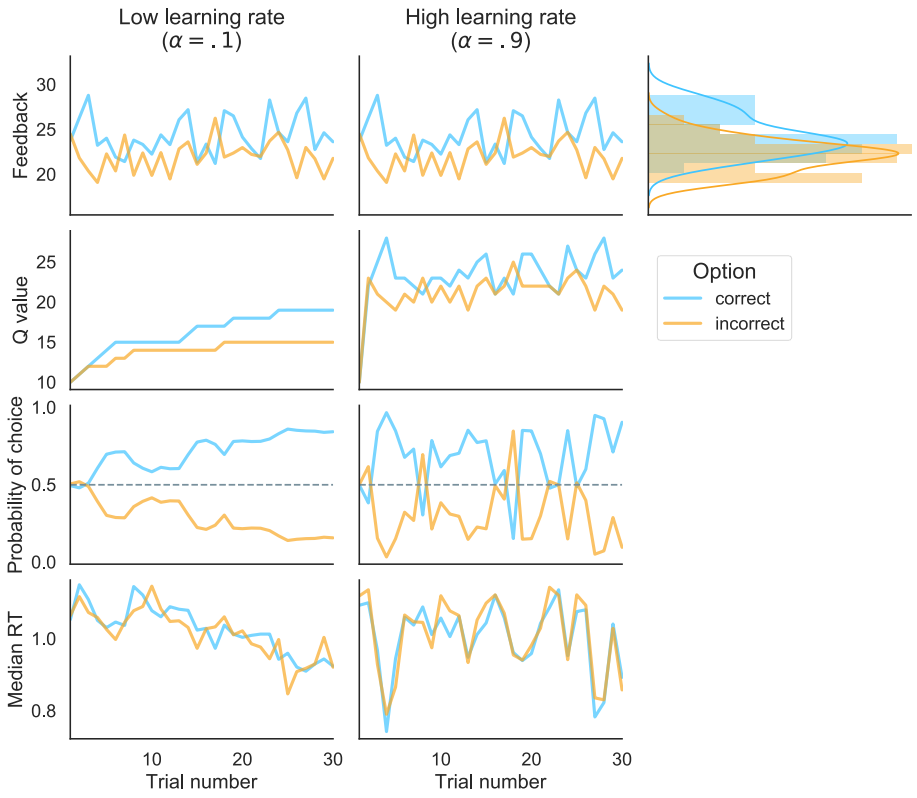


Figure 1.3: Reinforcement learning diffusion decision models describe both trial-by-trial and within-trial dynamics. As in Figure 1.2, the same feedback (first row) is given to two different agents (two columns), that update the expectations (Q values) with higher or lower learning rates. In each trial, decisions are similar to the process illustrated in Figure 1.1, where the drift-rate of evidence accumulation in proportional to the difference in Q values between the correct and incorrect options. While the probability of choosing the correct or incorrect option is similar to Figure 1.2, this model also makes predictions on the response times.

and valence on RTs, with participants being the slowest in the loss-partial condition. These effects did not depend on the particular experiment, although participants were generally faster in experiment 4 (given the higher time pressure) (see Figure 2.2).

Moreover, we wanted to test whether the effects of valence on RTs and the effects of feedback on both RTs and accuracy could be explained by latent learning variables of a previously proposed RL model for choice accuracy in this task, i.e., the RELATIVE model (Palminteri et al., 2015). The learning rule in this model is the Rescorla-Wagner rule (Equation 1.3), with separate learning rates for the chosen and unchosen options, and a separate learning rate for the learning context. The decision rule in this model was the softmax rule (Equation 1.4). To avoid fitting the same data twice, we fitted a Bayesian version of the RELATIVE model to the choice accuracy on a subset of the data and inferred the latent learning variables for new unseen data. In particular, we were interested in the absolute difference in learned values between the two options in a trial $|\Delta Q_t|$, and the learned context value V_t (i.e., that corresponds to the expected average outcome of a particular pair of options). We then ran two separate Bayesian mixed linear models on choices and RTs[5] and confirmed that, while $|\Delta Q_t|$ predicted choices and RTs, V_t only predicted RTs (Figure 2.3).

Both these analyses, however, were limited, since they could not explain correlations between choices and RTs. Therefore, we moved to the SSM framework and fitted the DDM to choice accuracy and RTs simultaneously. To test learning contexts effects on the DDM parameters across the four experiments, we fitted a three-layered hierarchical Bayesian DDM, where the bottom layer corresponds to the participants, the middle layer corresponds to the experiments, and the top layer corresponds to the whole dataset. To account for possible effects of context on the threshold, drift-rate, and non-decision time parameters of the DDM[6], we fitted separate intercepts and three coefficients per parameter (corresponding to valence, feedback information, and their interaction). Note that, in these analyses, we did not account for trial-by-trial effects due to learning, but only of the across-trial effects of different learning contexts. At a dataset level, we found that feedback information affected all three parameters: In partial contexts, the drift-rate and threshold were lower and the non-decision time was higher. On the other hand, valence affected the

[5] We controlled for general increase of accuracy and decrease of RTs by adding the number of trial as predictor.

[6] The bias in the starting point was not considered, as options were randomized to the left or right side of the screen

non-decision time, with higher non-decision time values in the loss domain. Valence also affected the threshold, with higher threshold values in the loss domain, but only in experiments 1, 2, and 3. There was also an interaction between valence and feedback on the threshold, with the lowest threshold in the reward-partial context (Figure 2.4).

While the feedback effect on the drift-rate is similar to a difficulty effect, the higher cautiousness in complete feedback contexts might be explained by a regret for not choosing the opposite option. The non-decision time was higher in the decision contexts that were the most disadvantageous: when feedback was partial and in the presence of losses. This effect is similar to an avoidant Pavlovian response causing a delay in the response. Importantly, slower RTs did not trade off with an increase in accuracy (i.e., the extra decision time did not contribute to evidence accumulation). However, the slowing down of participants in the loss domain was partially explained also by a higher threshold parameter (suggesting an additional effect of losses on cautiousness which trades-off with higher accuracy), although this effect was overall smaller than the non-decision time effect and disappeared in the experiment where there was a higher time pressure.

Overall, these results show how RTs are crucial for understanding human behavior during learning-by-feedback, as some of the learning context effects were not visible on choices alone. These effects were also robust across the four experiments. Interestingly, losses mostly had an effect on RTs (i.e., the effect on accuracy was not significant), contrary to previous accounts of choice behavior in the presence of losses in non-learning environments (Kahneman & Tversky, 1979). This effect, together with the slowing down effect of presenting partial feedback, was explained by a DDM parameter that is rarely considered in the SSM literature: the non-decision time. While a previous study in behavioral economics (Kocher & Sutter, 2006) has shown how time-dependent payoffs could affect behavior in a similar way (by decreasing RTs without loss of accuracy), previous accounts of punishment-avoidant behavior always predicted a slowing down effect together with an increase of accuracy.

Since the analyses in this manuscript were mainly done to capture context effects across learning, the context-independent learning effects, as the increase of accuracy and decrease in RTs throughout the trials, were not modeled nor explained by the DDM.

1.1.5 Second manuscript: A reinforcement learning diffusion decision model for value-based decisions

In the second manuscript (Chapter 3), I analyzed data from an RL experiment in which participants were presented with different learning contexts. The task was a *n*-armed bandit task. In each block, a total of four new options (identified by figures, see Figure 3.2) were presented, in groups of two at a time[7] , and participants had to choose between one of them within 3 seconds. The rewards underlying the four options were normally distributed, with same variance but different means, and were all in the gain domain (Figure 3.1). The pairs of options were chosen so that options could differ in their overall value (i.e., *magnitude*) and in the difficulty of choice. The pairs AB, AC, BD, and CD, had, respectively, overall mean value 38, 43, 47, 52. Plus, the value difference in pairs AB and CD was 4 on average (i.e., difficult trials), and the difference in pairs AC and BD was 14 on average (i.e., easy trials).

To test the effect of magnitude in the gain domain (from low to high), difficulty (easy vs. difficult), and of their interaction on accuracy and RTs, we fitted two separate Bayesian mixed linear models (see Figure B.1). While difficulty had the often reported effect on both accuracy and RTs (participants were slower and less accurate when choosing between AB and CD), magnitude only affected RTs, with participants being faster when choosing between higher-valued options pairs. There was also an interaction of difficulty and magnitude on the RTs, with participants being the slowest in the difficult, low value condition (i.e., pair AB) (Figure 3.3 and Figure 3.4).

We then estimated three classes of models: RL, DDM, and RLDDM, using hierarchical Bayesian parameter estimation. The models were compared both quantitatively, by means of the WAIC, and qualitatively, by means of posterior distributions of the mean accuracy and RTs across learning contexts (in order to check whether the models could capture the context effects) and binned trials in the learning session (in order to check whether the models could capture the effects of learning).

RL models varied in the learning rule, which was the Rescorla-Wagner rule (Equation 1.3) with either one learning rate, or separate learning rates for positive and negative RPE, as in Pedersen et al. (2017). The decision rule was the softmax rule (Equation 1.4)

[7] Differently from the previous study, the same option could appear in different pairs

with either fixed or increasing sensitivity parameter, as a function of the number of times an option is seen, as in, e.g., Yechiam and Busemeyer (2005). The RL model with separate learning rates for positive and negative RPE and fixed sensitivity described the data best, both qualitatively (Figure 3.5) and quantitatively (Table 3.1). Importantly, the RL model correctly described the learning curves of accuracy throughout learning and a higher accuracy when deciding between pairs AC and BD. However, the RL model did not make predictions for the RTs.

We then fitted three DDMs. In the first and simpler DDM, there were two drift-rates, for easy and difficult choices, one threshold, and one non-decision time. In the second DDM, there were two drift-rates (as in the first DDM) but four thresholds, corresponding to the four choice pairs, AB, AC, BD, and CD. The third and last DDM had four separate drift-rates and four separate thresholds, corresponding to the four choice pairs (as in the second DDM). The second and third DDMs explained the data equally well, both qualitatively (Figure 3.6) and quantitatively (Table 3.2). In particular, the DDM correctly described the difficulty effect on both accuracy and RTs, as well as the magnitude effect on RTs. The difficulty effect was driven by higher drift-rates in the easy compared to difficult choice pairs, while the magnitude effect was driven by lower threshold parameters in the higher – compared to lower – valued pairs. However, the DDMs could not predict the learning curves of either accuracy or RTs.

Finally, we fitted different combinations of RLDDMs: We considered different learning rules, as well as different mechanisms to explain the mapping of learned values to the DDM parameters in each trial. As in the RL models – and as in Pedersen et al. (2017) – the learning rule was the Rescorla-Wagner rule, with either a single learning rate or separate learning rates for positive and negative RPE. The threshold was either fixed or could be modulated by the context value (i.e., the mean of the presented options in a trial), as described in Equation 3.6. We proposed this mechanism in order to account for the magnitude effect on RTs, unlike in Pedersen et al. (2017). Finally, the mapping between the difference of values of the presented options and the drift-rate could be either linear – as in Pedersen et al. (2017) and Frank et al. (2015) – or a sigmoid function, as described in Equation 3.7. This mechanisms was also not considered in previous RLDDMs, and we included it to test whether the assumption that the drift-rate grows linearly with the value differences is valid across higher and lower difficulty levels. The model that explained the data best, both qualitatively (Figure 3.7) and quantitatively (Table 3.3), was a model with (1) separate learning rates for positive and negative RPE – as in the RL model comparison

and as found by Pedersen et al. (2017), (2) a context modulation of the threshold on a trial-by-trial base, with lower thresholds corresponding to contexts with higher values – as in the across-trial DDM analyses, and (3) a non-linear mapping function between value differences and drift-rate. Importantly, the preferred RLDDM could capture the context effects on both accuracy and RTs, as well as the learning curves of accuracy and RTs throughout the trials in a learning session.

In Pedersen et al. (2017)'s study, however, the RLDDMs also included different mechanisms, such as an increasing drift-rate across the trials following a power function, and a decreasing threshold across the trials, also following a power function. We therefore estimated four additional models, with separate learning rates for positive and negative RPE and different combinations of these mechanisms. The preferred model was – as in Pedersen et al. (2017) – a RLDDM with a power-decreasing threshold and without a power-increasing drift-rate. However, none of these models outperformed our preferred RLDDM, either quantitatively (Table 3.4) or qualitatively (Figure 3.8): Not only could these models not explain the magnitude effects on RTs, but they all underestimated accuracy in the difficult condition, and could not capture the decrease in RTs in the difficult condition throughout learning.

In sum, we proposed a novel RLDDM to account for the magnitude effect, as well as for different difficulty levels in a RL task. We also extended previous applications of RLDDMs to learning-by-feedback when rewards are normally distributed. We show how important it is to consider diverse learning context in order to challenge previous models' assumptions. First, previously proposed RLDDMs assumed that performance only depends on value differences and not on the overall value of the presented options. By providing participants with options with different overall values in the gain domain, we observe that they get faster when higher valued options are presented. This effect can be explained as a context, trial-by-trial modulation of the threshold, where the threshold decreases in the presence of higher valued pairs of options. This effect was found in a non-learning context by Cavanagh et al. (2014), and was interpreted as a striatal facilitation effect due to increased dopamine levels in the presence of higher rewards (Wiecki & Frank, 2013). Second, previously proposed models assumed that the value differences scaled linearly with the drift-rate in each trial. By providing participants with different levels of difficulty, we showed how models that assume linear mapping functions between the value difference and the drift-rate fail to account for both accuracy and RTs in the difficult trials (as they consistently underestimate performance in these trials). By assuming a sigmoid function

to map value differences to the drift-rate parameter, we were successful in predicting both accuracy and RTs across difficulty conditions.

1.2 The neural bases of decision under uncertainty

Computational modeling has proven to be a useful tool also in cognitive neuroscience. Just as cognitive psychologist, cognitive neuroscientists aim to study the processes underlying behavior. However, cognitive neuroscientists are also interested in how these processes are implemented in the brain, using proxies for neural activity, such as EEG, MEG, and fMRI data. By comparing neural data across experimental manipulations, one can infer the involvement of a particular brain area in a task. Cognitive models can help this inference process by making quantitative predictions on the latent processes that explain differences in behavior across conditions. Therefore, instead of correlating neural data and raw behavior, or of contrasting neural data across experimental condition, one can correlate neural data to a model's parameters. This is useful as it improves the psychological interpretation of a brain region's role in a task. On the other hand, by correlating neural data to computational models, these models can be potentially falsified if, e.g., one can prove that they make biologically implausible predictions. The mutual benefit of cognitive modeling and cognitive neuroscience is at the core of the field of model-based cognitive neuroscience (Forstmann & Wagenmakers, 2015).

In Section 1.1, I presented two classes of models, SSMs and RL models, and combinations between the two, RLDDMs. In Sections 1.2.1 and 1.2.2, I will give an overview on the literature in cognitive neuroscience that links these models to neural data. One of the most robust findings in the field of model-based cognitive neuroscience is the finding of a neural correlate of the RPE – as described in Equation 1.2 – in dopaminergic neurons in the brain. In Section 1.2.3, I will explain why most of the evidence in this field comes from invasive, animal studies and not from non-invasive, human studies. Finally, in Section 1.2.4, I will present work that clarifies the contradicting findings of previous humans studies with the help of ultra-high field MRI.

1.2.1 Sequential sampling and the brain

In the last decades, studies in cognitive neuroscience have supported the idea of evidence accumulation in the brain, as described by SSMs (for an overview, see Gold & Shadlen, 2007; Huk & Meister, 2012; Hanks & Summerfield, 2017; Mulder, VanMaanen, & Forstmann, 2014; Heitz, 2008). Using extracellular recordings from single neurons in cortical areas of macaque monkeys performing perceptual discrimination tasks, two main regions have found to mirror evidence accumulation: the lateral intraparietal cortex (LIP) and the frontal eye fields (FEFs). In these tasks, monkeys are typically trained to respond by making saccades to one or the other side of the visual field. After aggregating stimulus-locked firing rates across trials and across these areas, the signals are shown to increase when evidence in favor of a saccade in the neurons' receptive fields is presented. Moreover, the signal increases more or less when the evidence is less or more noisy. Finally, when looking at response-locked aggregate firing rates, signals reach the same level, independently of the noise in the evidence. More careful inspection of these patterns has revealed heterogeneity in the selectivity of the neurons in these areas, as well as across trials, and it is still unclear to some extend whether LIP and FEFs are necessary for evidence accumulation (Hanks, Ditterich, & Shadlen, 2006; Katz, Yates, Pillow, & Huk, 2016). Nonetheless, unilateral activation of these areas biases contralateral saccadic choices (e.g., Katz et al., 2016; Wilke, Kagan, & Andersen, 2012) thus supporting the involvement of these areas in perceptual decision making as formalized by the SSM framework.

While single neurons recordings are not feasible in human studies because of their invasiveness, other neuroimaging techniques such as magnetoencephalography (MEG), EEG, and fMRI have been used to investigate evidence accumulation in humans performing typically more complex tasks. Despite its relatively poor spatial resolution, EEG's high temporal resolution allows to inspect the within-trial dynamics of decision making. MEG and EEG studies have shown a potential very similar to the signal in LIP and FEFs, referred to as the centroparietal positive potential (CPP), in perceptual decision making (Kelly & O'Connell, 2013; O'Connell, Dockree, & Kelly, 2012) and value based decision making (Gluth, Rieskamp, & Büchel, 2013). While this signal is centralized, the lateralized preparatory motor signals also reflects evidence accumulation contralateral for a specific response (i.e., right and left) across domains (Kelly & O'Connell, 2013; O'Connell et al., 2012; Gluth et al., 2013; Donner, Siegel, Fries, & Engel, 2009). On the other hand, fMRI studies have focused more on trial-by-trial changes and on individual differences of SSM

parameters. Multiple studies in perceptual decision making using fMRI provided evidence for a relationship between the caudate nucleus and pre-SMA with the decision threshold (Bogacz, Wagenmakers, Forstmann, & Nieuwenhuis, 2010), both across (e.g. Forstmann et al., 2008; van Maanen, Fontanesi, Hawkins, & Forstmann, 2016; Forstmann et al., 2010) and within (e.g. van Maanen et al., 2011) participants. Other studies looked at trial-by-trial flactuations in the rate of evidence accumulation and areas related with commitment to a task (e.g., Turner, van Maanen, & Forstmann, 2015; Turner, Wang, & C.Merkle, 2017). However, by presenting evidence in a value based decision making task at a slower pace, a correlate of evidence accumulation was shown in pre-supplementary motor area (pre-SMA), caudate nucleus, and anterior insula while trial-by-trial fluctuations in the decision threshold were associated with pre-SMA and caudate nucleus activity (Gluth, Rieskamp, & Büchel, 2012).

1.2.2 Reinforcement learning and the brain

RL models have become increasingly popular in cognitive psychology in the 80s because they make quantitative predictions about the cognitive and neural mechanisms underlying learning-by-feedback. In the late 80s and 90s, thanks to pioneering work in electrophysiological recordings in behaving monkeys, a neural correlate of the RPE was found in dopamine neurons (for an overview, see, e.g. Dayan & Daw, 2008; Niv, 2009; Dayan & Abbott, 2001; Schultz, 2015; Watabe-Uchida, Eshel, & Uchida, 2017). Recent work in optogenetics has established a causal link between RPE-coding by dopaminergic neurons and learning (Steinberg et al., 2013). Dopamine neurons are mostly situated in the midbrain, and are concentrated in two regions, the ventral tegmental area (VTA) and the substantia nigra (SN), especially in one of its subdivisions, the pars compacta (SNc). The firing of neurons in these areas increases when rewards exceed expectations and decreases when rewards are less than expected. Crucially, the signal disappears when rewards are correctly predicted. Moreover, the neurons' firing is proportional to the RPE magnitude and it extends to cue stimuli associated with positive or negative RPE in previous trials. The current view on dopamine is that it represents subjective value (Schultz, 2010, 2015), by incorporating in the RPE variables such as the variance of the expected outcomes (i.e., risk) (Fiorillo, Tobler, & Schultz, 2003), and the time of the reward (Fiorillo, Newsome, & Schultz, 2008).

SN and VTA play a crucial role in the cortico-basal ganglia system, thus regu-

lating adaptive, goal-directed behavior (Holroyd & Coles, 2002; Haber & Knutson, 2010): By modulating synaptic activity in the ventral striatum (VS), dopamine facilitates actions towards rewards (i.e., approaching behavior). Berke (2018) has recently proposed a framework in which, by signalling to different striatal areas, dopamine neurons modulate the allocation of limited internal resources (i.e., energy for movements, attention, and time) based on the learned reward expectations. Moreover, Bromberg-Martin, Matsumoto, and Hikosaka (2010) suggested a possible distinction between two dopamine populations: one encoding motivational value (i.e., whether an outcome is better or worse than expected) and one encoding motivational salience (i.e., how surprising an outcome is, independently on whether is better or worse than expected) as well as a general alerting signal. While the first is mainly situated in VTA and projects to VS, which in turns sends projections to the ventro-medial prefrontal cortex (vmPFC), the second is mainly situated in SNc and projects to dorsal striatum (DS), which in turns projects to dorso-lateral prefrontal cortex (dlPFC) (Matsumoto & Hikosaka, 2009). Both functions are crucial for adaptive behavior: Motivational value promotes actions that maximize rewards, while motivational salience promotes learning in the presence of changes in stimulus associability (Pearce & Hall, 1980).

1.2.3 The dopamine reward signal in the human brain: Challenges and previous findings

Activity from the SN and VTA has been mostly studied in electrophysiological studies with animals. This is due to a series of methodological challenges that arise when measuring the SN and the VTA signal with non-invasive imaging techniques such as fMRI.

First of all, the midbrain is situated deep in the brain and far from the skull. The further away an area is from the receive elements of the scanner, the lower the signal-to-noise ratio is. At the same time, the SN and the VTA are quite small (around 511 mm3 and 138 mm3, respectively, see Table 4.1) and they neighbor each other, as well as other nuclei such as the red nucleus and the subthalamic nucleus. The VTA is also close to the cerebrospinal fluid, which constitutes a source of physiological noise. Therefore, not only the midbrain signal is low, but the chances of mixing up the signal coming from different small midbrain nuclei are very high. This has been previously reported as the "subcortical cocktail problem" (de Hollander, Keuken, & Forstmann, 2015). This problem can be exacerbated by common MRI procedures. For example, by increasing the spatial resolution of fMRI images to get a more detail view on the midbrain area, the signal-to-noise ratio further

decreases. On top of this, spatial smoothing – commonly used in fMRI analyses to increase signal-to-noise ratio and to correct for disalignment between the individual space and the standard space [8] – increases the chance of including the signal from neighboring areas.

Another challenge has to do with the chemical composition of SN: The SN has a high concentration of iron, and, thus, different magnetic properties from, e.g., cortical tissue. In particular, the T_2^* signal in iron-rich areas, which is what is measured in functional images, has a faster decay. Because fMRI protocols are usually tailored to cortical and subcortical areas with much lower iron concentration, the protocols used in standard practices are suboptimal when measuring signal in the SN.

In order to study the signal of the subthalamic nucleus, a midbrain nucleus with similar iron concentration to the SN, de Hollander, Keuken, van der Zwaag, Forstmann, and Trampel (2017) proposed a protocol for 7 Tesla MRI. First, thanks to a stronger magnetic field, 7 Tesla MRI provides increased signal-to-noise and contrast-to-noise ratio at higher spatial resolutions (Eapen, Zald, Gatenby, Ding, & Gore, 2011). Second, shorter echo times allowed the measurement the T_2^* signal before it is completely decayed[9] . Finally, de Hollander et al. (2017) drew masks, individually for each subject and nucleus, in order to more reliably extract the signal from the areas of interest. By doing so, they could avoid using spatial smoothing and ensure that the signal from neighboring areas was not mixed. In this protocol, a fairly high resolution of 1.5 mm isotropic was obtained, and the temporal signal-to-noise ratio (tSNR) – which is a measure of quality of the fMRI time series – of this protocol was higher compared to different protocols using either 3 Tesla or 7 Tesla MRI.

With the exception of the study by Zaghloul et al. (2009), in which they measured activation of the SN using microelectrode recording during deep brain stimulation, previous studies with human subjects did not achieve such high resolution when measuring activity in the VTA and SN. D'Ardenne, McClure, Nystrom, and Cohen (2008), Pauli et al. (2015), and Zhang, Larcher, Misic, and Dagher (2017), measured activity of dopaminergic nuclei using 3 Tesla MRI. In the three studies, individual masks were not drawn in order to

[8] This is particularly important when trying to correctly localize a particular area in a subject's brain using standard space coordinates.

[9] Note that, in 7 Tesla MRI, the difference of T_2^* decay between different areas is higher than in 3 Tesla MRI. Therefore, shorter echo times are particularly important with 7 Tesla when acquiring functional images. Shorter echo times ask for faster acquisition, which leaves less time for artifact correction techniques, such as fat suppression. On the other hand, the higher T_2^* contrast in 7 Tesla compared to 3 Tesla is an advantage in structural images, because it allows to better delineate iron-rich nuclei from the neighboring areas.

carefully distinguish activation of neighboring nuclei, and they all used spatial smoothing: in D'Ardenne et al. (2008) using a 3 mm FWHM Gaussian kernel, in Pauli et al. (2015) using a 2 mm FWHM Gaussian kernel, and in Zhang et al. (2017) using a 4 mm FWHM Gaussian kernel. Although we do not have tSNR measurements for these studies, it is likely that they could not achieve high tSNR values, based on the results of de Hollander et al. (2017).[10]

These studies also provided partially contradicting results. D'Ardenne et al. (2008) only found a positive (and not negative) RPE in the VTA and no signal in the SN. Pauli et al. (2015) focused on the SN and found a positive (and not negative) RPE in the ventromedial SN and a negative (and not positive) RPE in the dorsolateral SN, together with a negative expected value (EV) signal. Finally, Zhang et al. (2017) also focused on the SN and found a RPE in the medial SN and a surprise signal in the lateral SN.

Therefore, to the best of our knowledge, previous studies measuring signal in the dopaminergic nuclei in human subjects using fMRI have not found common agreement with the results from animal studies. Moreover, while correlates of risk and surprise were found in cortical areas such as the anterior insula (AI) and the amygdala in the human brain, no study has yet measured such signals in the midbrain. These signals were so far only shown in animal studies (e.g., Fiorillo et al., 2003; Matsumoto & Hikosaka, 2009).

1.2.4 Third manuscript: The role of dopaminergic nuclei in predicting and experiencing gains and losses: A 7T human fMRI study

In the third manuscript (Chapter 4), we collected data from 27 human subjects in two separate sessions: one to collect structural and one to collect functional MRI data. During the structural session, a multi-echo magnetization-prepared rapid gradient echo (ME-MP2RAGE) sequence (Caan et al., 2018) was used to acquire perfectly aligned multimodal high-resolution (0.7 mm isotropic) images: T_1-weighted, T_2^*-weighted, and Quantitative Susceptibility Mapping (QSM; Langkammer et al., 2012) images. These images highlight different tissue contrast (Figure 4.2) that are necessary in order to delineate masks on a subject level: While SN is seen at best in QSM images – as they highlight differences in

[10] In this study, however, they replicated the protocol used by Pauli et al. (2015) finding lower tSNR in the midbrain nuclei.

magnetic properties of the tissues, VTA is seen at best when combining T_1-weighted and T_2^*-weighted images – as they highlight, respectively, the VTA border with the CSF and the VTA border with the SN and red nucleus. Right and left VTA and SN masks were drawn for each subject by two independent, trained raters. In order to measure the raters' agreement we computed the Dice score, which is the ratio of the intersection and the union of the masks: Scores equal to one correspond to perfect agreement, while scores equal to 0 correspond to no agreement. The agreement was higher in SN compared to VTA, likely because of its clearer borders and because it is a larger region (Table 4.1). In general, the agreement was higher than previously reported scores (Keuken & Forstmann, 2015) and, as a conservative measure, only the intersections of the masks across raters were kept for the subsequent functional analyses.

Moreover, when using masks defined in the standard space to extract the signal of VTA and SN in the individual space (such as the ones proposed by Keuken & Forstmann, 2015; Pauli, Nili, & Tyszka, 2018) there is the risk of including too little voxels from the area of interest as well as voxels from neighboring regions – a problem known as misalignment. To quantify the overlap between the neighboring structures using this method, we calculated Dice scores between the individual VTA and SN masks that we defined in the individual space, and previously proposed standard masks of SN and VTA subdivisions. Note that this measure does not include additional noise coming from the use of eventual spatial smoothing. We found the highest overlap between the medial part of the SN and the VTA (Figure 4.3), which is explained by the fact that the medial SN is the part neighboring VTA. Our results thus support the importance of drawing individual masks. However, drawing individual masks also requires significantly more work and high-quality multi-modal structural images, which are not always available in MRI studies because of time limitations (our structural sequences were acquired in about 20 minutes). In our case this was only possible because participants were invited in two separate sessions.

During the functional session, participants engaged in a gambling task. The task we used was an adaptation of the task by Preuschoff, Bossaerts, and Quartz (2006), and was chosen because it allowed us to (1) temporally separate the expectation from the delivery of gains and losses, (2) measure both the EV and risk before gains and losses are delivered, (3) measure both RPE and surprise (i.e., salience of the outcome) when gains and losses are delivered, and, finally, (4) it does not allow for excessive individual variability in the learning process, since participants are instructed about the reward structure of the task. Note that, despite being measured at the same point during the trial, risk and EV, as well

as RPE and surprise, are not correlated, thus allowing reliable parameter estimates in the general linear model analyses (and therefore to separate the associated signals).

In this task (Figure 4.1), two numbers between 1 and 5 are drawn without replacement in each trial. At the beginning of the trial, participants have to bet whether the second number will be higher or lower than the first. After a period of 4 to 10 seconds, the first number is shown and the EV and risk can be measured: if, e.g., a participant bets that the second number is lower and the first number is 2, then the EV is negative, with some variability (risk) in the possible outcomes (i.e., there is one chance on four to still win the bet). After another period of 4 to 10 seconds, the second number is shown, together with the corresponding gain of 5 euros (if they were correct) or loss of 5 euros (if they were incorrect). Since the bets are blind, the expected reward of a particular choice (i.e., "second number is lower", and "second number is higher") was the same across the experiment: $EV = 5 \cdot .5 - 5 \cdot .5 = 0$, and it was not possible to learn which action was most advantageous. At the same time, the task allowed us to measure signal correlated with the reward variables that we were interested in.

The main analysis of the fMRI data was done by averaging the signal across each region of interest (ROI) thus obtaining a time series for each ROI, block of trials (there were two in total), and participant (Figure 4.4). By running a GLM on the time series (with a correction for temporal autocorrelation in the signal), we found a significant correlation with the RPE (positive and negative, as described in Equation 1.2) in both the VTA and the SN, and no EV signal (positive and negative, defined as the mean expected reward in a trial). These results confirmed previous results from the animal literature (Schultz, 2015) and clarified previous fMRI results in humans, which did not consistently find a full RPE signal in VTA/SN and found a negative EV signal in SN (Pauli et al., 2015). Moreover, we showed – for the first time in human subjects – a risk signal (defined as the expected variance in the outcomes in a trial) in both the VTA and the SN and a surprise signal (defined as the unsigned RPE) in the SN alone. The presence of a risk signal is in line with previous findings in the animal literature (Fiorillo et al., 2003), and with the notion that dopamine represents deviations from the expected subjective value, by incorporating both the mean and variance in the reward expectations (Schultz, 2015). The presence of a surprise signal in SN alone, on the other hand, is in line with the framework proposed by Bromberg-Martin et al. (2010) in which there are two main dopamine populations, with separate functions for learning-by-feedback.

Finally, we performed the same GLM analyses at a voxel level in the rest of the acquired brain (Figure 4.5). This was not the entire brain: In order to acquire the desired spatial resolution and echo time, a few brain slices had to be sacrificed. These analyses were performed as a control, to check whether we could replicate previous findings in the human literature regarding neural correlates of reward variables outside the midbrain. Overall, we confirmed previous findings, by observing EV signal in orbital frontal cortex and anterior insula, risk signal in anterior cingulate cortex, amygdala, anterior insula, and dorsal striatum, RPE signal in ventral striatum, and anterior insula, and a surprise signal in the posterior insula. These results confirm that the experimental manipulations were successful in eliciting the expected neural responses to gains and losses.

Overall, these results showed that, thanks to recent advances in MRI methods, it was possible to measure a reward signal in the dopaminergic nuclei of behaving humans, without recurring to invasive techniques. Such signal is in line with current theories coming from studies with animals. This study opens the way to new research employing potentially more complex learning tasks, leading to a better understanding of the circuit connecting dopamine, striatum, and prefrontal areas.

1.3 Discussion

1.3.1 Modeling behavior in different learning contexts

In the first two manuscripts of this thesis, I showed how, by combining computational models from the perceptual (i.e., SSMs) and the economic (i.e., RL models) decision making traditions, we can better understand the learning and decision processes at play during learning-by-feedback (Summerfield & Tsetsos, 2012). Until recently, RL models only described choice data and, specifically, how choice preferences evolve with experience as a function of the received feedback. However, in RL models, no assumptions are made about how the decision variable evolves, from the presentation of the stimuli to the the selection of an option. As a consequence, RL models made no predictions on RT data. On the other hand, SSMs describe single decisions as accumulation-to-bound processes and make joint predictions about choices and RTs. Moreover, both models have been supported by neural data. In particular, accumulation-to-bound processes have been found in human and animal studies across the perceptual and the economic domains, and signal similar to

the updating of reward expectations, as described by RL models, has been found in the animal and human brain. In both cases, a causal relationship between neural processes and their relative functions has also been established.

In the first manuscript, we used the DDM (a widely used SSM) to explain participants' behavior across different learning contexts: when learning to choose the most advantageous of two losing or two winning options, and when partial or complete feedback is presented. In the loss domain, participants were consistently slower but not more or less accurate. On the other hand, when partial as opposed to complete feedback was provided, participants were consistently slower and less accurate. The DDM provided a mechanistic explanation of these effects: (1) Behavioral changes in the loss domain were mainly due to a higher non-decision time. The threshold was also higher in the loss domain, but not consistently across the experiments. (2) Behavioral changes in partial feedback contexts were due to a lower evidence accumulation rate, a higher threshold, as well as a longer non-decision time. However, these analyses were limited, as we only explained choice preferences and RTs across trials in the different learning contexts, without explaining how performance evolves throughout experience as a function of feedback. In order to do so, a tighter link between the RL variables and the DDM parameters has to be established.

The second manuscript goes in the direction of establishing such a link, by proposing a new combination of RL and SSMs (RLDDM) which builds on seminal work in this field (Pedersen et al., 2017; Frank et al., 2015). The proposed RLDDM was tested on a task in which different pairs of options could have higher or lower overall value and high or low value difference. This was done to challenge some of the assumptions of previous RLDDMs: (1) the assumption that only value difference influences behavior and not overall value, and (2) the assumption that the evidence accumulation rate is linearly proportional to value differences. Our data falsified both assumptions, and we proposed a model that could accommodate the observed behavioral patterns by means of: (1) a trial-by-trial threshold modulation mechanism, in which overall value of a context decreases the threshold in a trial, and (2) a sigmoid-shaped mapping between value differences and the evidence accumulation rate. Models that did not incorporate these mechanisms failed to predict the slowing-down effect in the presence of lower valued pairs of options, and underestimated accuracy and the decrease of RTs throughout learning when very similar options were presented.

In sum, both manuscripts highlight the importance of considering RTs during learning-by-feedback as an additional proxy for the underlying learning and decision pro-

cesses. At the same time, providing participants with different learning contexts – by manipulating difficulty, valence, and reward magnitude – allowed to observe interesting behavioral patterns.

1.3.2 Valence and magnitude effects

In the first two manuscripts, we observed a similar behavioral pattern in the presence of losses and when lower-value options were presented: participants got slower without getting more or less accurate. However, different accounts of these effects were given in the two manuscripts. In the first study, the loss-effect was explained by a higher non-decision time and – in experiments with less time pressure – a higher threshold. In the second manuscript, a trial-by-trial modulation in the threshold explained the magnitude effect, with higher thresholds when low-value pairs were presented. These two parameters – the threshold and the non-decision time – have different psychological interpretations: While a higher threshold corresponds to a higher cautiousness, a higher non-decision time reflects a general halt in the accumulation process, due to slower stimulus encoding or to motor inhibition.

While the valence effect has been mainly studied in learning paradigms in value-based decisions, the magnitude effect has been mainly studied in non-learning paradigms in both value-based and perceptual decisions.

Notably, Ratcliff and Frank (2012) combined a biologically plausible neural network model (Frank, 2006b) to the DDM to analyze behavioral data in a learning paradigm with both gains and losses. As in the first manuscript of this dissertation, participants were slower but not more accurate in the loss domain. Such effect was explained through the DDM[11] by either a higher threshold, a higher non-decision time, or by a threshold that collapses within a trial. In the neural model, activity in the loss-difficult context was caused by (1) the subthalamic nucleus, which inhibits the thalamus in the presence of conflict (i.e., in difficult trials) causing a general halt in the responses (i.e., the hyperdirect pathway of the basal ganglia circuit), as well as by (2) dopamine activity, which inhibits the thalamus through the striatum in the presence of losses, causing a slowing down in the response (i.e., the indirect pathway of the basal ganglia circuit). Finally, these two mechanisms interact

[11] The DDM did not capture learning effects, as our analyses in the first manuscript.

with each other, causing particularly slow responses in difficult-loss contexts. Although this study proposed a biologically plausible explanation for the valence and difficulty effects, it is still unclear which of the DDM parameters is mainly affected by losses: Is it the non-decision time or the threshold? And, since the indirect pathway has been more associated with threshold modulations, this model consistently predicts higher accuracy in the presence of losses (contrary to our and previous findings). A later study using the same paradigm has explained the valence effect by means of changes in the decision threshold (Cavanagh et al., 2014), although non-decision time effects were not tested.

Recently, Ratcliff, Voskuilen, and Teodorescu (2018) gave a DDM account of the magnitude effect in a series of perceptual decision making tasks. They observed that participants were faster when discriminating between two overall more intense stimuli. The effect of stimulus intensity on accuracy, however, varied across tasks, with some tasks showing no effect on accuracy, and others showing lower accuracy in the presence of more intense stimuli. This effect was captured by two different models, that explained the data equally well. The first model explained the magnitude effect as a change in the across-trial variability in the drift-rate, and the second model explained it as a change in the within-trial noise of the diffusion process[12] . Finally, based on theoretical reasons, they proposed the first model as an explanation of the magnitude effect across domains. On the other hand, Polania et al. (2014) directly compared the magnitude effect in perceptual and in value-based decisions: While magnitude did not affect either accuracy or RTs in the perceptual domain, in the value domain higher-valued options made participants faster and more accurate. In sum, it is to some extent unclear whether the magnitude effect is stable across domains and tasks, and which DDM parameter would capture the effects on on both accuracy and RTs best.

As shown in Figure 1.4, a higher drift-rate variability, as well as higher within-trial noise, lead to a strong decrease in accuracy and only a mild decrease in the RTs. On the other hand, lower threshold vales lead to a strong decrease in RTs and a non-linear decrease in accuracy. Only a lower non-decision time predicts faster responses without affecting accuracy. Therefore, future work should establish whether the valence and magnitude effects have little or no impact on accuracy altogether. In the case of the magnitude effect, this could very well depend on the domain (i.e., perceptual or economic) and on the specific task, thus reflecting different mechanisms at play. In case accuracy is not affected, then the non-decision time seems to be the better mechanism and might be linked to a general "hold

[12] Higher within-trial noise corresponds to decreased threshold and drift-rate.

your horses" response (i.e., hyperdirect pathway). In case accuracy is mildly affected, then the threshold would be a better mechanisms, and might be linked to dopamine activity (i.e., the direct and indirect pathways of the basal ganglia circuit). This could be further clarified by combing computational modeling with neural data. This work should build on recent findings showing separate EEG signatures for valence and magnitude (Yeung & Sanfey, 2004), as well as interactions between magnitude, valence and ambiguity on EEG signals (Gu et al., 2017).

1.3.3 Difficulty and feedback information effects

The first two manuscripts also proposed different accounts of the difficulty and partial feedback effects: While the difficulty effect was only explained by changes in the drift-rate (second manuscript), the partial feedback effect was explained by changes in drift-rate, threshold, and non-decision time. However, based on raw data alone, the two effects were very similar, consisting of a decrease in accuracy and increase in RTs.

On the one hand, these results suggest that receiving partial feedback might induce an avoidant response, similar to the one induced by losses. As discussed in the previous paragraph, it is an open question for future studies whether such a response is linked to a "hold your horses" mechanism in the basal ganglia circuit. On the other hand, the results of the second manuscript are in line with traditional accounts of the difficulty effect, which only predict changes in the drift-rate across domains (Ratcliff & Rouder, 1998; Milosavljevic et al., 2010). However, contrary to previous studies (Cavanagh et al., 2014), difficulty did not have an effect on the threshold. In line with the predictions of the neural network model proposed by Frank (2006b), however, both feedback information and difficulty had an interaction on the RTs with, respectively, valence and magnitude.

Finally, our RLDDM did not directly account for uncertainty in the outcomes. Based on elecrophysiological studies and on the results presented in the third manuscript, showing that risk and surprise are represented in the dopamine signal, additional mechanisms might need to be incorporated in the RLDDM to account for performance in learning environments with, e.g., more or less outcome variance, or in dynamic environments.

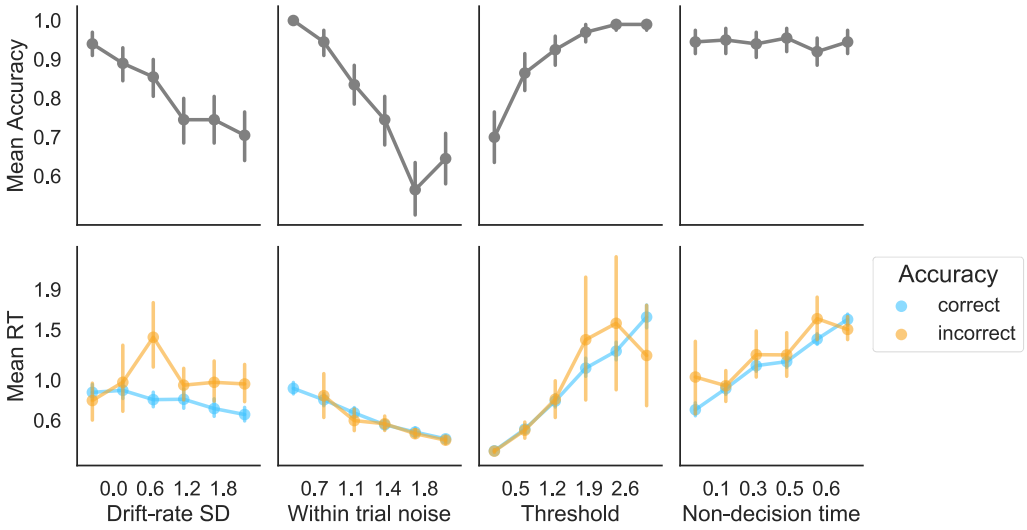


Figure 1.4: Changes in the diffusion decision model parameters differently affect choice preferences and response times (RTs). A higher across-trial noise in the drift-rate (first column) decreases accuracy and increases RTs in incorrect responses. A higher within trial noise (second column) decreases both accuracy and RTs. A higher threshold (third column) causes a non-linear increase in the accuracy, as well as a linear increase in the RTs. A higher non-decision time (fourth column) causes a linear increase in the RTs, and does not affect accuracy. 200 trials were simulated for each parameter combination. Bars represent 95% confidence intervals.

1.3.4 The dopamine signal

In the third manuscript, we showed how, by capitalizing on high quality structural and functional images acquired with 7 Tesla MRI, it was possible to measure a reward signal in the dopaminergic nuclei – the VTA and the SN – in behaving humans. To date, studies using fMRI in humans provided mixed findings regarding this signal, and did not investigate whether risk was also represented in these areas.

First of all, we showed the importance of carefully delineating the VTA and the SN at a participant level. This is because they are relatively small and close to each other, as well as to other nuclei with different functions, such as the subthalamic nucleus (e.g., Frank, 2006a). Previous studies (e.g., D'Ardenne et al., 2008; Zhang et al., 2017; Pauli et al., 2015) not only used lower quality images (3 Tesla MRI), but they used population-based coordinates to identify the VTA or the SN (or subdivisions of the SN), and spatial smoothing, thus likely mixing the signals of neighboring nuclei and obtaining low temporal signal to noise ratio values.

By carefully delineating these regions at a participant level, we found a clear RPE signal and an absence of a EV signal in both the SN and the VTA. This is in line with findings from animal studies (Schultz, 2015): The firing of dopaminergic neurons increases when rewards exceed previous expectations, decreases when rewards are less than expected, and does not change when rewards meet previous expectations. We also found a risk signal in both the VTA and the SN. This is in line with previous electrophysiological studies in monkeys: The firing of dopaminergic neurons was found to vary with reward probability, in the presence of stimuli anticipating rewards (Fiorillo et al., 2003). Finally, we found a surprise signal in the SN and not in the VTA. This result supports the recent framework proposed by Bromberg-Martin et al. (2010). In this framework, two sub-populations of dopamine neurons, one in the SN and one in the VTA, fire for motivational value and motivational salience, respectively. Therefore, the two populations facilitate two complementary aspects of learning: Learning to approach a rewarding option and avoid a punishing option, and learning about the environment's volatility. In order to further validate this hypothesis, future work should extend the current findings to more complex learning tasks, e.g., to dynamic learning environments or by setting a higher cost to failed predictions.

Building on these results, and to further validate this framework, the connectivity between the VTA and the SN and different areas of the striatum should be investigated.

Finally, based on neural models of the basal ganglia circuit (Frank, 2006b), neural data can be linked to RLDDM models, to further clarify the differential function of dopamine neurons and of the subthalamic nucleus in regulating decisions during learning.

1.3.5 Conclusion

According to Marr (1982)'s terminology, computational models can be described on three different levels: computational, algorithmic, and implementational. The computational level defines the objective of a decision problem: in the case of RL, the objective is reward maximization and punishment minimization; in the case of SSM, the objective is the optimization of both RTs and accuracy. The algorithmic level defines how this objective is met. Specific instances of RL and SSM describe optimal ways in which these objectives are met (Niv, 2009; Sutton & Barto, 1998; Bogacz et al., 2006). Finally, at the implementational level, there is evidence pointing to populations of neurons computing similar algorithms to the ones described by both RL models and SSMs: While neurons in the dopaminergic nuclei code for RPE, neurons in associative cortical areas show activity correlated with bounded evidence accumulation (see Section 1.2.1 and 1.2.2). However, evidence from electrophysiological studies suggests that the firing of neurons in these areas is highly heterogeneous, both functionally and temporally (e.g., Hanks & Summerfield, 2017; Watabe-Uchida et al., 2017). To what extent the implementation level should be tightly linked to the algorithmic level (i.e., at the neuron level, or at a region level) remains an open question in the field.

Computational models constitute an important tool for the understanding of adaptive and goal-directed behavior in healthy as well as in clinical populations. They can be improved by advancements at the algorithmic level, by extending them to explain behavior in different domains and tasks. They can also be improved at the implementational level, as neural data can help to clarify disputes between models that are equally good at a behavioral level but make different predictions on the neural mechanisms involved.

In this thesis, I propose a new model that combines aspects of evidence accumulation models and RL (computational and algorithmic level), and show how process data – such as RTs – can offer new insight on learning and decision processes. Furthermore, I provided evidence for a reward signal in the human midbrain (implementation level), capitalizing on novel MRI methods.

Chapter 2

Decomposing the Effects of Context Valence and Feedback Information on Speed and Accuracy During Reinforcement Learning: A Meta-Analytical Approach Using Diffusion Decision Modeling

Laura Fontanesi, Maël Lebreton[1], and Stefano Palminteri[1]

The manuscript is under revision in *Cognitive, Affective, and Behavioral Neuroscience*.

Financial disclosure: SP is supported by an ATIP-Avenir grant (R16069JS) Collaborative Research in Computational Neuroscience ANR-NSF grant (ANR-16-NEUC-0004), the Programme Emergence(s) de la Ville de Paris, and the Fondation Fyssen. ML is supported by an NWO Veni Fellowship (Grant 451-15-015). The Institut dEtudes de la Cognition is supported financially by the LabEx IEC (ANR-10-LABX-0087 IEC) and the IDEX PSL* (ANR-10-IDEX-0001-02 PSL*). LF is supported by grants 100014_153616/1 and P1BSP1_172017 from the Swiss National Science Foundation. The funding agencies did not influence the content of the manuscript.

[1] Co-supervision

Abstract: Reinforcement learning (RL) models describe how humans and animals learn by trial-and-error to select actions that maximize rewards and minimize punishments. Traditional RL models focus exclusively on choices, thereby ignoring the interactions between choice preference and response time (RT), or how these interactions are influenced by contextual factors. However, in the field of perceptual decision making, such interactions have proven to be important to dissociate between different underlying cognitive processes. Here, we analyzed such interactions in behavioral data from four RL experiments, which feature manipulations of two factors: outcome valence (gains vs. losses) and feedback information (partial vs. complete feedback). A Bayesian meta-analysis revealed that these contextual factors differently affect RTs and accuracy: While valence only affects RTs, feedback information affects both RTs and accuracy. To dissociate between the latent cognitive processes, we jointly fitted choices and RTs across all experiments with a Bayesian, hierarchical diffusion decision model (DDM). The drift-rate parameter was uniquely affected by the feedback manipulation, with a higher drift-rate in complete feedback conditions, similarly to difficulty effects. Moreover, there was an interaction effect on the threshold, with lowest thresholds in the reward-partial condition, indicating a possible effect of regret. Finally, the non-decision time was affected by both manipulations, with lower non-decision times in the most advantageous learning contexts (in gain domains and with full feedback), suggesting a possible motor facilitation in these contexts. These results showed how, by explaining RTs and choice data during RL using the DDM, we can gain a better understanding of the mechanisms underlying decisions in different learning contexts.

2.1 Introduction

In cognitive psychology, the sequential sampling modeling (SSM) framework has enabled the development of models which jointly account for choice accuracy and response time (RT) data in two-alternative forced choice tasks (Gold & Shadlen, 2007; Bogacz et al., 2006; Smith & Ratcliff, 2004; Ratcliff & Smith, 2004). In this framework, it is assumed that, when evaluating two choice options, evidence in favor of one over the other alternative(s) is accumulated over time and a response is initiated when this evidence reaches a decision threshold. The crucial advantage of applying these models to empirical data is that they can help decompose the interactions between RTs and accuracy into meaningful psychological concepts. On the one hand, speed and accuracy can be positively correlated: e.g., when faced with easy decisions, people tend to give more correct and faster responses compared to when facing difficult decisions (Ratcliff & Rouder, 1998). This effect is captured in SSMs by higher rates of evidence accumulation. On the other hand, speed and accuracy can also be negatively correlated: e.g., when asked to make speedy decisions, people tend to be less accurate (Ratcliff & Rouder, 1998). This phenomenon is referred to as the speed-accuracy tradeoff (Heitz, 2008; Luce, 1986) and is explained within the SSM framework by a decrease in the decision threshold and interpreted as reduced cautiousness. Finally, speed and accuracy can also be uncorrelated: e.g., people can differ in how fast or slow they respond, without being more or less accurate (Ratcliff, Thapar, & Mckoon, 2003). These differences are captured in SSMs by the non-decision time parameter, which represents motor processes necessary for the execution of actions as well as time needed for stimulus encoding. Therefore, SSMs have provided a mechanistic explanation of these three different correlation patterns of RTs and accuracy and have been successfully applied in various psychological domains: from perceptual, to social, to economic decision making, as well as in memory and language research (Ratcliff et al., 2016).

Research in reinforcement learning (RL) aims at characterizing the processes through which agents learn, by trial-and-error, to select actions that maximize the occurrence of rewards and minimize the occurrence of punishments (Sutton & Barto, 1998). A century-long experimental investigation of RL processes in human and non-human animals has shown that learning is accompanied by a simultaneous increase of the frequency of the selection of the most advantageous action and by a decrease of the time necessary to select this action (Pavlov, 1927; Skinner, 1938; Thorndike, 1911).

However, traditional computational RL models only account for choices and do

not consider RTs (but see the recent work of Frank et al., 2015; Pedersen et al., 2017). Therefore, how contextual factors in RL paradigms impact the relation between RTs and accuracy is still relatively poorly understood (Summerfield & Tsetsos, 2012).

In a series of recent studies, Palminteri and colleagues (Palminteri et al., 2015; Palminteri, Kilford, Coricelli, & Blakemore, 2016; Palminteri, Lefebvre, Kilford, & Blakemore, 2017) developed an RL paradigm where they orthogonally manipulated two important contextual factors: feedback information and outcome valence. Feedback information was modulated by showing (i.e., complete feedback) or not showing (i.e., partial feedback) the outcome associated with the unchosen option. Outcome valence was modulated by reversing the sign of the outcome (i.e., gains vs. losses), which directly impacted the goal of learning: reward-seeking vs. punishment-avoidance. Independent analyses reported in the aforementioned studies consistently show that: First, participants learned equally well to seek rewards as to avoid punishments; second, participants displayed a higher accuracy in complete feedback contexts. Importantly, RTs in the same task follow a different pattern: Participants were slower in the punishments contexts and in partial-feedback contexts.

By looking through the lenses of SSMs, different decision processes may drive the behavioral patterns reported in these studies. In the present paper, we first re-assess the effects of the contextual factors on RTs and accuracy using a meta-analytical approach involving data from four behavioral experiments employing the same RL paradigm. Then, we fit a previously proposed RL model (Palminteri et al., 2015) to look at relationships between latent learning variables, estimated in a subset of the data, and the remaining raw behavioral data. We found that while the learned contextual value (i.e., the overall value of a pair of choice options) only predicted RTs, the difference in learned values predicted both accuracy and RTs. Finally, we moved to the SSM framework: We used a hierarchical Bayesian version of the standard diffusion decision model (DDM, Ratcliff, 1978) to test the effects of the contextual factors (i.e., feedback information and valence) on the model's parameters (i.e., drift-rate, threshold, and non-decision time) across the four experiments. We found that the rate of evidence accumulation was higher in full feedback compared to partial feedback contexts, cautiousness was the lowest in the gain domain when the feedback information was partial, and the non-decision time increased in the loss domain as well as when the feedback was partial. Altogether, our results illustrate that accounting for RTs in instrumental learning paradigms provides valuable information about the decision processes underlying learning by feedback.

2.2 Methods

2.2.1 Participants

We analyzed data from four behavioral experiments, realized in three different research centers in France and UK (final N=89; Table 2.1). The local ethical committees approved the studies and participants provided written informed consent; see the original publications for additional details (Palminteri et al., 2015; Salvador et al., 2017).

2.2.2 Task

Participants performed a probabilistic instrumental learning task designed to manipulate both feedback valence (reward vs. punishment) and feedback information (partial vs. complete) using a 2x2 factorial design (Figure 2.1 A). Participants had to choose one of two abstract cues (letters from the agathodaimon font). Each trial (Figure 2.1 B) started with a fixation cross, followed by presentation of the cues during which participants indicated their choice. After the choice window (either 3 or 1.5 seconds, depending on the experiment), a red arrow highlighted the chosen option. Then, the outcome was revealed, and participants moved to the following trial. In each session, there were eight different cues, divided into four fixed pairs, corresponding to four choice contexts: reward-partial, reward-complete, punishment-partial, and punishment-complete. In reward contexts, the best cue had 75% probability of yielding a reward (points or money) and 25% probability of yielding nothing; while the worst cue, on the other hand, had 25% probability of yielding a reward and 75% probability of yielding nothing. In punishment contexts, the best cue had 25% probability of yielding a loss and 75% probability of yielding nothing, while the worst cue had 75% probability of yielding a loss and 25% probability of yielding nothing. In partial feedback contexts, participants were presented with only the outcome of the chosen cue, while in complete feedback contexts they were presented with the outcomes of both the chosen and forgone cues. The number of trials per context, the number of sessions, and the timing slightly differed across experiments (see Table 2.1).

Table 2.1: Participants.

	Experiment 1	Experiment 2	Experiment 3	Experiment 4
Sample size	20	25	20	24
Mean age	25.4	23.9	32.4	22.2
Percentage Males	55	36	55	38
Response window (sec)	3	3	3	1.5
N sessions	2	3	3	2
N trials per session	80	96	96	80
Center	Paris - ENS	Paris - ENS	Paris - ICM	London- UCL
Source	Pilot for	Pilot for	Controls	Controls
Reference	(Palminteri et al., 2015)	(Palminteri et al., 2015)	(Salvador et al., 2017)	(Palminteri et al., 2016)

Note. Demographics, task characteristics, and investigation centers of the four experiments (N: sample size, ENS: cole Normale Suprieure; ICM: Institut du Cerveau et de la Molle; UCL: University College London).

2.2.3 Dependent variables

Our main dependent variables were the correct choice rate (accuracy) and RTs. A correct response is defined as a choice directed toward the best (reward maximizing or punishment minimizing) cue of a pair. The RT is defined as the time between the presentation of the options and the button press. In order to include only trials whose cumulative accuracy was overall higher than 50%, the first 12 trials of each session were discarded in the ANOVA and DDM analyses (Figure 2.1 C-D), but not in the RL modeling, in which the first trials are crucial, and linear mixed-effect analyses, where trial number was explicitly entered as a predictor.

2.2.4 Bayesian analysis of the variance

Accuracy and RTs were analyzed in two independent ANOVAs, which modeled the main effects of – and the interaction between – the experimental manipulations (i.e., valence and feedback information). We adopted a Bayesian mixed model meta-analysis approach, where the different experiments could be modelled as fixed effects (Singmann et al., 2014). By doing so, we could test whether, across the four experiments, mean accuracy and RTs differed and whether the learning contexts were similar across the experiments.

This approach entails a comparison of different Bayesian models using Bayes Factors (BFs) (Kass & Raftery, 1995; Wagenmakers, 2007) in a two-step procedure. First, we

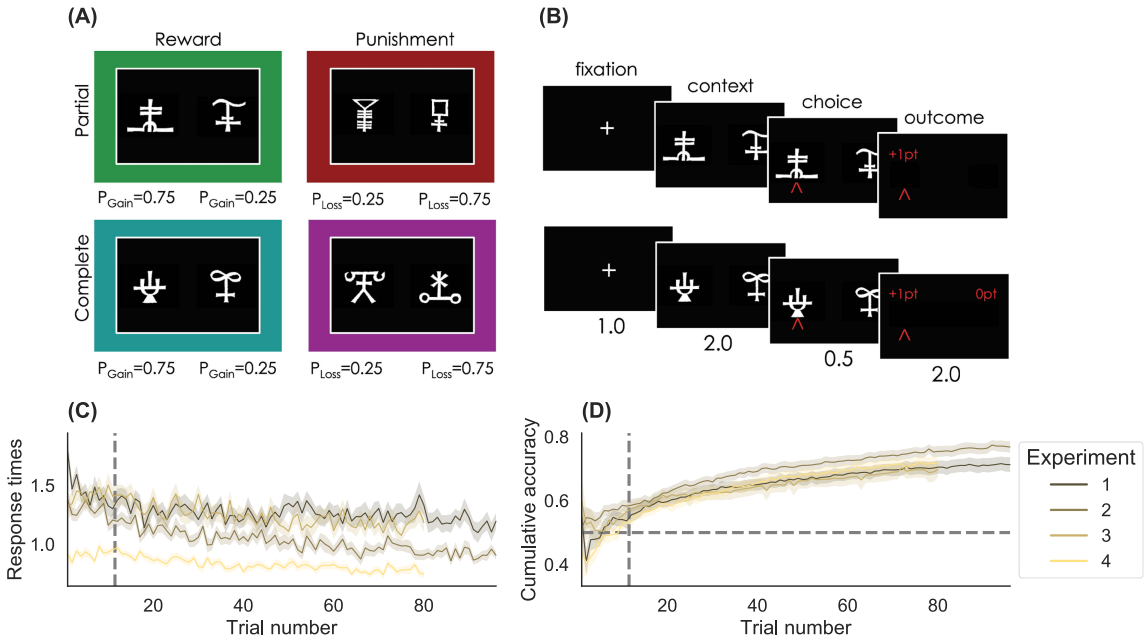


Figure 2.1: *Task factors and learning curves.* (A) The learning task 2x2 factorial design. Different symbols were used as cues in each context, and symbol to context attribution was randomized across participants. The coloured frames are purely illustrative and represent each of the four context conditions throughout all figures. "Reward": gain domain; "Punishment": loss domain; "Partial": only feedback of the chosen option is provided; "Complete": both feedback of chosen and unchosen options are provided; P_Gain= probability of gaining 1 point; P_Loss= probability of losing 1 point. (B) Time course of example trials in the reward-partial (top) and reward-complete (bottom) conditions. Stimuli durations are given in seconds. (C) Average response times during learning. (D) Cumulative accuracy during learning. Trials before the vertical dotted lines were discarded for the across-trial analyses. The horizontal dotted line in (D) indicates chance level.

assessed if the experiments should be treated as fixed effects by comparing such a model to a model with only the random effect of participants. The winning model was then used as a baseline model in the second step, where we assessed which combinations of fixed-effects and interactions gave the most parsimonious, but complete account of the data. Once we identified the best model, we inspected the estimated posterior distribution of its main effects and interactions (see Appendix A.1). The models were all fit using the R package BayesFactor (Morey, Rouder, & Jamil, 2015) and adapted code previously provided by Singmann et al. (2014).

2.2.5 Reinforcement learning architecture

To capture the trial-by-trial dynamics due to learning-by-feedback, we fitted the "RELATIVE" model, proposed by Palminteri et al. (2015). This model is based on a simple Q-learning model (Sutton & Barto, 1998), but allows separate learning rate parameters for outcomes of chosen and forgone options, and includes a contextual module, so that option values are updated relative to the learned value of the choice context.

In the RELATIVE model, at each trial t, the option values Q in the current context s are updated with the Rescorla-Wagner rule (Rescorla and Wagner, 1972):

$$Q_{c,s,t} = Q_{c,s,t-1} + \alpha_c \cdot \delta_c$$
$$Q_{u,s,t} = Q_{u,s,t-1} + \alpha_u \cdot \delta_u$$

where α_c is the learning-rate for the chosen option Q_c – updated in both partial and complete feedback contexts – and α_u the learning-rate for the unchosen option Q_u – updated only in complete feedback contexts. δ_c and δ_u are prediction error terms, calculated as follows:

$$\delta_c = R_{c,s,t} - V_{s,t-1} - Q_{c,s,t-1}$$
$$\delta_u = R_{u,s,t} - V_{s,t-1} - Q_{u,s,t-1}$$

V_s represents the context value that is used as the reference point for the updating of option values in a particular context, and R is the feedback received in a trial. Context value is also learned via a delta rule:

$$V_{s,t} = V_{s,t-1} + \alpha_V \cdot \delta_V$$

where α_V is the learning-rate of context value and δ_V is a prediction error term. In complete feedback contexts:

$$\delta_V = \frac{(R_{c,s,t} + R_{u,s,t})}{2} - V_{s,t-1}$$

In partial feedback contexts, since $R_{c,s,t}$ is not provided, its value is replaced by its expected value $Q_{u,s,t}$, hence:

$$\delta_V = \frac{(R_{c,s,t} + Q_{u,s,t})}{2} - V_{s,t-1}$$

The decision rule was implemented as a softmax function:

$$p_{A,s,t} = \frac{e^{\theta Q_A}}{(e^{\theta Q_A} + e^{\theta Q_B})}$$

where p_A is the probability of choosing an option A over an option B and θ is the sensitivity parameter.

2.2.6 Reinforcement learning model fitting

To avoid testing the same data twice, we fitted a hierarchical Bayesian version of the RELATIVE model on the choice data of experiment 1 (training set), and generated predictions for the data of experiments 2, 3, and 4 (testing set). To ensure generalizability of the results, the same procedure was then repeated using data of experiments 2, 3, 4 as training sets instead (see Appendix A.2). To generate predictions for the testing set, we sampled individual parameters from the group-level parameter distributions estimated on the training set. This set of individual parameters (i.e., predictions for new, unseen, participants)[2] were then used to predict the latent variables of the RELATIVE model on a trial-by-trial base, given the feedback received by the participants in the testing set.

The RL model was coded and fitted using *stan*, a probabilistic programming language for Bayesian parameter estimation (Carpenter et al., 2017). The learning-rate parameters were given the following prior distributions:

$$\mu_\alpha \sim \mathcal{N}(.8, .5)$$
$$\sigma_\alpha \sim \mathcal{HN}(0, .5)$$
$$\alpha \sim \phi(\mathcal{N}(\mu_\alpha, \sigma_\alpha))$$

where μ_α is the group-level mean, σ_α is the group-level standard deviation, and α is the individual learning-rate. $\mathcal{N}$ is the normal distribution (with parameters mean and standard deviation), $\mathcal{HN}$ is the half-normal distribution, and ϕ is the cumulative density function

[2] See Figure A.1 for the individual parameters distributions used to generate predictions.

of the standard normal distribution, transforming α so that $0 \leq \alpha \leq 1$. The sensitivity parameter was given the following prior distribution:

$$\mu_\theta \sim \mathcal{N}(-1, 1)$$
$$\sigma_\theta \sim \mathcal{HN}(0, .5)$$
$$\theta \sim exp(\mathcal{N}(\mu_\theta, \sigma_\theta))$$

where μ_θ is the group-level mean, σ_θ is the group-level standard deviation, and θ is the individual sensitivity, which was exponentially transformed, so that $\theta \geq 0$. To estimate the joint posterior distribution of the model, we ran 4 independent chains with 5000 samples each, and discarded the first half of each chain. To test for convergence, we checked that the $\hat{R}$ statistic (Gelman & Rubin, 1992) – a measure of convergence across chains – was lower than 1.01 for all parameters.

2.2.7 Relationship between latent learning variables and raw data

At each trial, the choice difficulty (captured by the unsigned difference in options expected values $|\Delta Q_t|$), the trial contextual value V_t, and the trial number, were used as independent predictor variables in two linear regression models, respectively modeling accuracy and RTs. In these analyses, both participants and experiments were treated as random effects. For completeness, we reported Bayes Factors of the full model, and competing reduced models in Appendix A.2. The regression models were ran using the *BayesFactor* (Morey et al., 2015) package in R.

2.2.8 Diffusion decision model architecture

The DDM (Ratcliff, 1978; Ratcliff & Rouder, 1998) assumes that, when deciding between two alternatives, evidence in favor of one relative to the other is accumulated in time, according to the following differential equation:

$$dx = \mathcal{N}(v \cdot dt, c \cdot \sqrt{dt}), x_0 = a/2 \tag{2.1}$$

where dx is the change in the accumulated evidence in the time interval dt, v is the mean accumulated evidence across the time intervals, and c is the noise constant, usually fixed

to 1 [3] . A decision is executed when enough relative evidence in favor of an alternative has been collected, which is when x is either lower than 0 or higher than the decision threshold a. When the decision is unbiased (i.e., there is equal initial evidence in favor of both options), then the evidence accumulation starts from half the threshold a. In the experiments that were considered in the present study, the upper boundary corresponded to the correct option (i.e., the option with the highest mean payoff) and the lower boundary corresponded to the incorrect option (i.e., the option with the lowest mean payoff) within a context. Because these options were randomly assigned to the right and left sides of the screen, we assumed that decisions were always unbiased, and coded responses as correct and incorrect.

Therefore, the execution time and probability of choosing the option with the highest payoff depended on three main parameters. The first is the decision threshold a: Lower thresholds lead to faster but less accurate decisions, while higher thresholds lead to slower but more accurate decisions. The threshold is usually interpreted as response caution, with higher thresholds corresponding to higher cautiousness. The second parameter is the drift-rate v, which is the amount of evidence accumulated per unit of time. This can reflect the difficulty of the decision problem, as well as participants' efficiency in the task: Higher drift-rates lead to faster as well as more accurate responses. The third parameter that we take into account is referred to as *non-decision time* (NDT), and reflects the processes that influence the decision time, but do not pertain to evidence accumulation per se, such as motor and stimuli encoding processes. The non-decision time therefore affects RTs without affecting accuracy.

2.2.9 Diffusion decision model fitting

For each of the DDM parameter (i.e., v, a, and NDT), we fitted an intercept and three slopes, corresponding to the two main effects – valence and feedback information – and their interaction. This allowed us to test the effects of the experimental manipulations on the model parameters. To account for all levels of variability, we used a three-level version of the hierarchical Bayesian DDM, where the first level corresponds to the participants, the second corresponds to the experiments, and the third corresponds to the whole dataset,

[3] This is done to be able to identify the other parameters. One could decide to fix a different parameter, e.g., the decision threshold, to estimate this variable instead.

thus mimicking the meta-analysis approach described in Section 2.2.4.

The following prior distributions were assumed for the parameter intercepts:

$$v_{\text{int}} \sim t_{10}(3,1) + z_i + z_j$$
$$a_{\text{int}} \sim Gamma(1,1) + z_i + z_j$$
$$NDT_{\text{int}} \sim U(0,1)) + z_i + z_j$$

where t_{10} is the Student t distribution with 10 degrees of freedom and parameters mean and standard deviation, *Gamma* is the gamma distribution with parameters shape and scale, and U is the uniform distribution with lower and upper-boundaries as parameters. The following prior distributions were assumed for the parameter coefficients (coefficients corresponding to the main and interactions effect were given the same priors):

$$v_{\text{coeff}} \sim t_1(0,5) + z_i + z_j$$
$$a_{\text{coeff}} \sim t_1(0,5) + z_i + z_j$$
$$NDT_{\text{coeff}} \sim t_1(0,5) + z_i + z_j$$

z_i and z_j respectively account for individual ($1 \leq i \leq 89$) and experiment ($1 \leq j \leq 4$) deviations from the group mean: z_i represents the deviation of a participant's parameter from that parameter mean in the experiment, while z_j represents the deviation of an experiment's parameter mean from the parameter means across the overall dataset. The following prior distributions were given to z_i and z_j:

$$z_i \sim \mathcal{N}(0, \sigma_j)$$
$$z_j \sim \mathcal{N}(0, \sigma)$$

σ_j ($1 \leq j \leq 4$) and σ respectively account for the within- and across-experiment variances, and have priors:

$$\sigma_j \sim t_{10}(3,0)$$
$$\sigma \sim t_{10}(3,0)$$

To fit the Bayesian DDM and estimate its joint posterior distribution, we used *brms* (Bürkner, 2017), an R package for Hierarchical Bayesian model fitting based on *stan* (Carpenter et al., 2017). We first ran the model separately by experiment in order to get plausible starting values, using 5000 samples per chain and 2 chains, and discarding the first 2000 samples in each chain. We then ran the full model (across experiments) using the

same number of samples and chains. To test for convergence, we checked that the $\hat{R}$ statistic was less than 1.01, for all parameters, as in the RL analyses. To test the reliability of the parameter estimates, we performed parameter recovery on a simulated dataset (Palminteri, Wyart, & Koechlin, 2017); see Appendix A.4.

Finally, to assess the model fit of the DDM, we computed the posterior predictive distributions (Gelman, Meng, & Stern, 1996) for mean accuracy and RTs, as well as for RT quantiles (separately for correct and incorrect responses; Figure A.5).

2.2.10 Statistical reporting

In all analyses (i.e., ANOVA, linear mixed-effect regression and DDM), we report the estimated Bayesian credible interval (BCI) of the posterior distributions of the parameters of interest, computed as the 95% central interval of the distributions.

In all analyses, valence was coded as 0 for reward and 1 for punishment, and feedback was coded as 0 for partial and 1 for complete. Intercepts therefore correspond to the reward-partial context. The interaction was obtained by multiplying valence and feedback.

2.3 Results

2.3.1 Bayesian analysis of the variance

We assessed the effects of outcome valence and feedback information on learning performance (i.e., mean accuracy and RTs, Figure 2.2 A), using a Bayesian mixed model meta-analysis approach (see Section 2.2).

For the accuracy, our approach favored a model with (1) a single main effect accounting for feedback information, (2) no main effect of the experiment, (3) no interactions between experiment and experimental manipulations (M3 in Table A.1). These results indicates that only feedback and not valence had an effect on accuracy, and that this effect had a similar size across the experiments. The model parameters confirmed that accuracy was higher in the complete feedback information contexts ($\text{BCI}_{\text{Feedback}} = [.03 - .06]$) (see

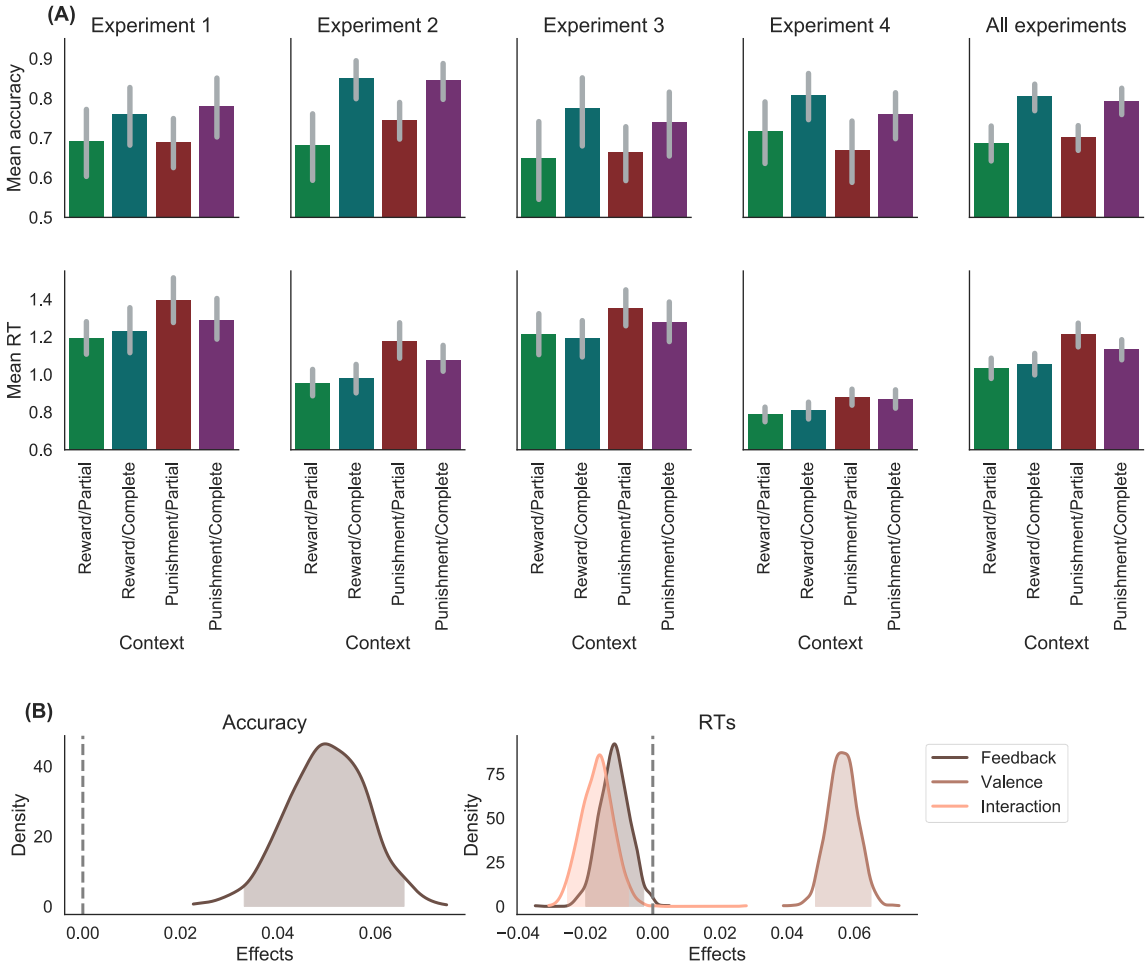


Figure 2.2: *Performance and behavioral effects across learning.* (A) Summary of the behavioral performance. Mean accuracy (top) and response times in seconds (bottom) are plotted, separately for experiments and conditions, as well as across experiments (right column). The bars represent 95% confidence intervals. (B) Posterior distributions of the feedback, valence, and feedback-valence interaction effects on accuracy and RTs of the preferred models in the ANOVA model comparison analyses. Shaded areas represent the 95% Bayesian Credible Intervals.

Figure 2.2 B).

For the RTs, our approach favored a model which includes (1) both main effects of valence and feedback information as well as their interaction, (2) a main effect of the experiment, (3) and no interaction between experiment and experimental manipulations (M5 in Table A.2). These results indicate that both valence and feedback information, as well as their interaction, had an effect on RTs, in a similar way across the experiments. The main effect of the experiment indicates that participants had different mean RTs across the experiments. The model parameters revealed that participants were slower in the loss domain ($BCI_{Valence}$ = [.05 – .07]) and faster in the complete feedback contexts ($BCI_{Feedback}$ = [-.019 – -.003]). In addition, the effect of valence was weaker in the complete feedback contexts ($BCI_{Interaction}$ = [-.03 – -.01]).

2.3.2 Reinforcement learning model analyses

As a first step toward understanding the processes underlying the effects of different learning contexts – outlined by the ANOVA – on both accuracy and RTs, we fit a RL model and inspected the relationship between latent learning variables and raw data. In particular, we fitted the RELATIVE model proposed by Palminteri et al. (2015) to a training set (i.e., data from experiment 1) (see Figure A.1), and generated predictions for a testing set (i.e., experiments 2, 3, and 4)[4] .

In particular, we were interested in predicting two latent variables of the RELATIVE model, as they develop in the testing datasets during learning. The first variable is the unsigned difference between the available option values $|\Delta Q_t|$, and the second one is the context value V_t. On the one side, $|\Delta Q_t|$ reflects choice difficulty, with lower values corresponding to more difficult choices. Throughout the trials, choices become easier, particularly in the complete compared to the partial feedback contexts (Figure 2.3 A). On the other side, V_t reflects valence: it increases in the reward contexts, and decreases in the punishment contexts, the more so in complete compared to partial contexts (Figure 2.3 A). In addition, we considered a third control variable, corresponding to the number of trial within learning session, to account for learning- and feedback-independent changes in the

[4] Note that the same procedure was repeated using data of experiments 2, 3, 4 as training sets instead, yielding similar results (see Figure A.3).

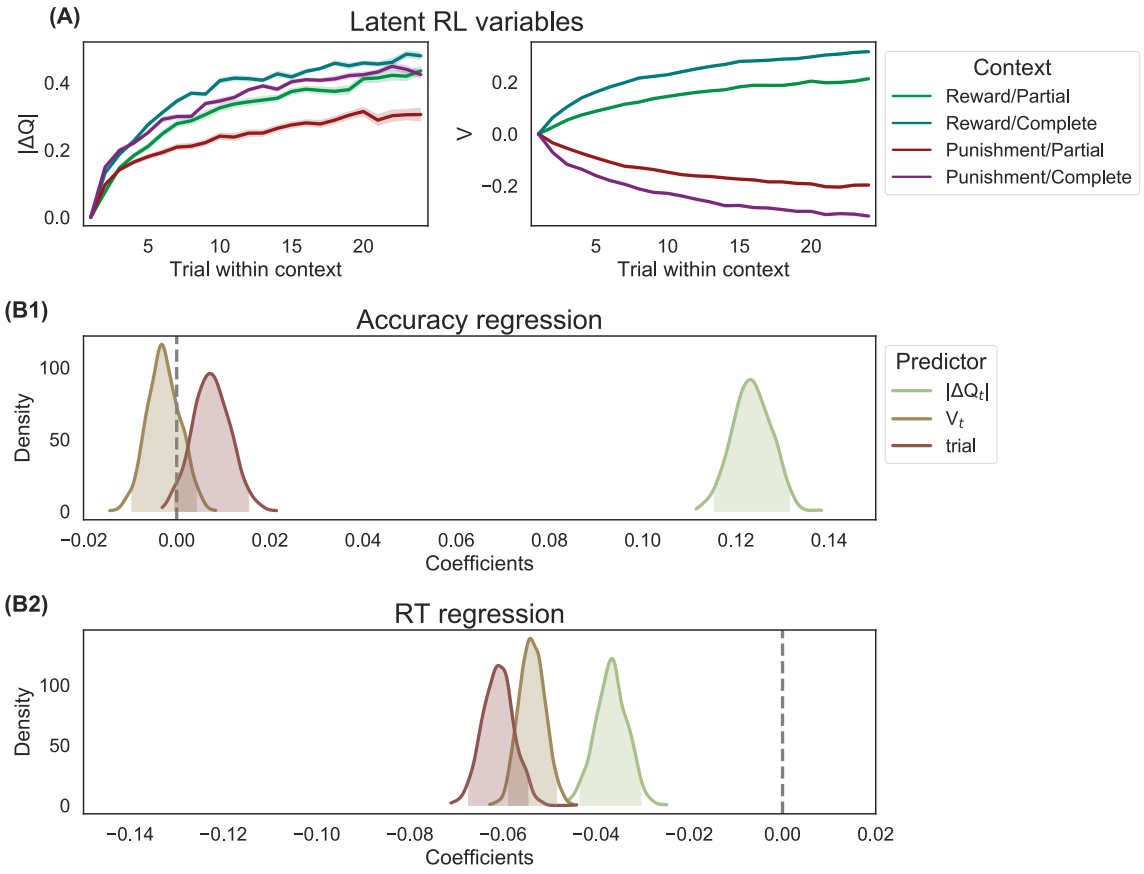


Figure 2.3: *Relationship between learning variables and performance.* (A) Predicted latent variables of the RELATIVE model, which are the unsigned difference in Q values $|\Delta Q_t|$ and the context value V_t. $|\Delta Q_t|$ is higher in complete information contexts compared to partial contexts. V_t is positive in the gain domain and negative in the loss domain, and is learned more quickly in complete contexts, where more information is available. Shaded areas represent 95% CI around the posterior mean. Predictions were made for experiments 2, 3, and 4, based on the data of experiment 1. (B1 and B2) Estimated posterior distributions of the linear models coefficients corresponding to the $|\Delta Q_t|$, V, and trial predictors. Shaded areas here represent the 95% BCI.

RTs and accuracy within the learning sessions.

The predicted $|\Delta Q_t|$ and V_t, together with trial number, were used as independent variables in Bayesian linear models of accuracy and RTs (Figure 2.3 B; but see also Figure A.2). We found that $|\Delta Q_t|$ was a good predictor of accuracy and RTs ($BCI_{accuracy}$ = [.12 – .13]; BCI_{RTs} = [-.04 – -.03]). On the other hand, V_t exhibited a different pattern, affecting only RTs, but not accuracy ($BCI_{accuracy}$ = [-.009 – .004]; BCI_{RTs} = [-.06 – -.05]; see also Table A.2 for the model comparison results). These results confirm the general trend that was described in the ANOVA results on a trial-by-trial base (based on a computational model of learning) and with out-of-sample predictions: The learned context value affects participants' speed of responses while providing complete feedback makes learning easier for the participants.

2.3.3 Diffusion decision model analyses

Although both the ANOVA and reinforcement learning analyses depict a consistent picture of the effect of different learning contexts on both RTs and accuracy, they do not model the interactions between the two. To decompose the simultaneous effects of contextual effects on RTs and accuracy, we therefore fitted a three-level hierarchical Bayesian version of the DDM to the data of all four experiments (see Section 2.2).

The increase in accuracy and speed in the complete feedback contexts was captured by an effect on all three DDM parameters (Figure 2.4): Providing participants with complete feedback increased the drift-rate (BCI = [.04 – .82]), increased the threshold (BCI = [.03 – .31]), and decreased the non-decision time (BCI = [-.073 – .003]). Compared to the gain domain, decisions in the loss domain showed higher non-decision time (BCI = [-.001 – .067]). Valence did not significantly affect the drift-rate (BCI = [-.50 – .23]) nor the threshold (BCI = [-.15 – .4]). However, the valence effect on the threshold was different across the four experiments, with stronger effects in experiments 1, 2, and 3, and the weaker effect in experiment 4 (Figure A.4). This might be due to the higher time pressure in experiment 4. Yet, we found a negative interaction between feedback information and valence on the threshold (BCI = [-.27 – -.05]). A closer examination of the threshold parameter by context (Figure 2.4, right column) revealed that the threshold was particularly low in the reward-partial condition. There was also a mild positive interaction effect on the non-decision time (BCI = [-.03 – .7]), indicating that the feedback effect on non-decision time was higher in

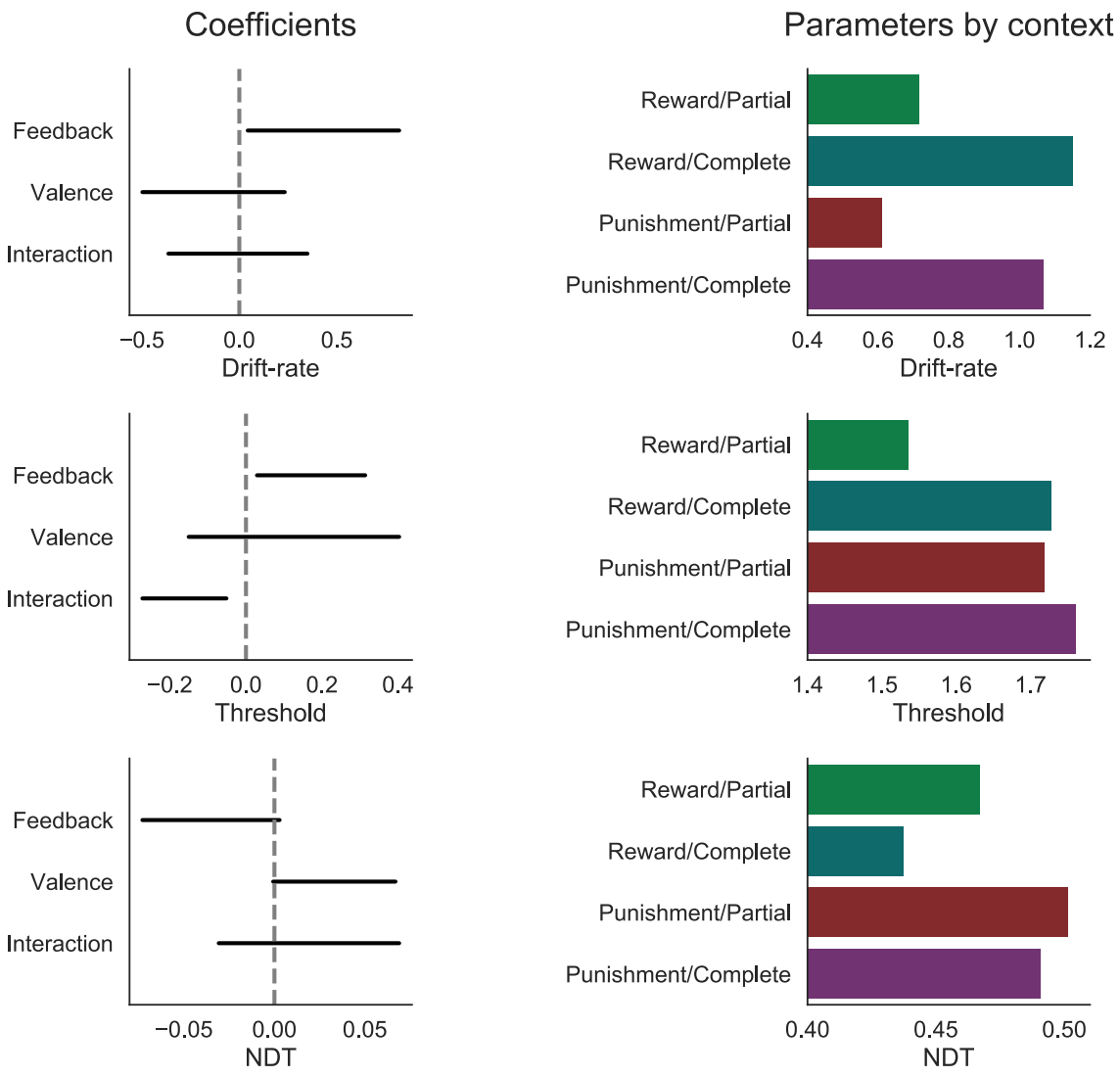


Figure 2.4: *Estimated diffusion decision model (DDM) parameters.* Left column: 95% Bayesian Credible Intervals of the estimated posterior distributions of the effects of the experimental manipulations (i.e., feedback information, outcome valence and their interaction) on the DDM parameter coefficients at the dataset level. Right column: estimated mean parameters at the dataset level, separately by context.

the gain domain. There was, however, no interaction effect on the drift-rate (BCI = [-.36 – .35]). In Figure A.4 we report the posterior distributions of the group parameters separately for experiments and for the overall dataset.

2.4 Discussion

In the present study, we looked at how different RL contexts (i.e., partial vs full feedback, and gains vs losses) affect accuracy and RTs. To do so, we used different methods and a relatively large dataset, composed of four separate experiments carried out in different centers.

First, we used a meta-analytic Bayesian approach to the analysis of variance of accuracy and RTs. Replicating previous reports (Palminteri et al., 2015, 2016; Salvador et al., 2017), we showed that participants were slower in the loss (as compared to the gain) domain, but not more or less accurate, and that they were more accurate and faster when complete (as compared to partial) feedback was provided. Interestingly, the similar accuracy observed in the gain and loss domains is at odds with the notion of loss aversion (Kahneman & Tversky, 1979): If in our task "losses loomed greater than gains", we would expect higher accuracy in the loss domain. However, by inspecting the RTs, we found that losses made participants slower, showing the importance of considering performance as a whole.

These results were further supported by the RL analyses: According to the RELATIVE model (Palminteri et al., 2015) predictions, the learned context value affects RTs on a trial-by-trial base and above learning. In this model, context value is used as reference point in a particular context to update the Q values in each trial. Palminteri et al. (2015) showed that including context value in the RELATIVE model improves the model fit to choice data (by comparing the RELATIVE model to similar RL models without contextual learning). Here we showed that context value can also be used to explain RTs data. Because the RELATIVE model decision rule (i.e., the sotfmax rule) does not allow it to predict RTs, this relationship has not been investigated so far. Moreover, within the RELATIVE model, we can quantify choice difficulty in each trial (or decision conflict, see, e.g., Cavanagh et al. (2014)) as the unsigned difference of the learned value of the available options. In line with previous studies that investigated the difficulty effect in value-based decision making on both accuracy and RTs (e.g., Milosavljevic et al., 2010; Cavanagh et al., 2014; Frank et

al., 2015; Lebreton, Jorge, Michel, Thirion, & Pessiglione, 2009; Shenhav, Straccia, Cohen, & Botvinick, 2014), we show that choice difficulty is a good predictor for out-of-sample accuracy and RTs data.

Previous studies that applied the SSM framework to value-based decision making have shown how the difficulty effect can be captured by a decrease in the mean accumulation rate (Milosavljevic et al., 2010; Cavanagh et al., 2014; Frank et al., 2015; Krajbich et al., 2010). However, previous studies investigating the valence effect have given mixed interpretations (Ratcliff & Frank, 2012; Cavanagh et al., 2014). Because the reported RL analyses do not allow to inspect RTs and accuracy simultaneously, and to better understand this effect on RTs (as well as the interaction between valence and feedback information on RTs), we turned to the SSM framework and fitted the DDM simultaneously to accuracy and RTs across the four experiments. The effect of feedback information (i.e., higher accuracy and speed in the complete contexts) appeared to be driven by an increase of the drift-rate and of the threshold parameters, and by a decrease of the non-decision time in the complete compared to partial conditions. On the other hand, valence had a main effect on the non-decision time (with higher non-decision time in the loss domain), and there was an interaction of feedback and valence on the threshold (with lowest threshold in the reward-partial condition). The effect of valence on threshold (higher thresholds in the loss domain) was not consistent across experiments, and it was higher in experiments with less time pressure.

While drift-rate difficulty effects have been documented in both economic and perceptual decision making (Milosavljevic et al., 2010; Krajbich et al., 2010; Ratcliff & Rouder, 1998), the decrease in threshold in partial feedback contexts may appear counter-intuitive at first glance, as less information, and therefore higher uncertainty, could increase cautiousness. Moreover, previous studies have found that higher difficulty also leads to an increase in the threshold (e.g., Frank et al., 2015). Yet, a possible psychological interpretation for this effect is that the outcomes corresponding to the unchosen options are known to elicit regret, which can increase cautiousness in decision making (Zeelenberg, 1999; Shenhav et al., 2014). This can thus explain the interaction effect on the threshold, since regret should be the lowest in the reward-partial condition. This hypothesis could be further supported by a trial-by-trial mechanism, in which higher feedback of the unchosen options in complete feedback contexts causes an increase in the threshold in the following trial.

The two effects on the non-decision time (of both feedback and valence) are

less clear, as non-decision time effects are not very common in the SSM literature, and are thought to reflect stimulus encoding or purely motor processes (Ratcliff & Rouder, 1998). However, previous work in behavioral economics has shown that, by providing time-dependent payoffs under high time pressure, RTs can be reduced without any loss of accuracy (Kocher & Sutter, 2006). Nonetheless, Ratcliff and Frank (2012) found that models with increased threshold or increased non-decision time explained equally well the effect of losses on RTs. They further linked such effect to the dopamine modulation in the basal ganglia circuit (i.e., the indirect pathway). In a later study, however, (Cavanagh et al., 2014) explained the effect of losses on RTs with a threshold effect. Alternative accounts of the RT slowing in the loss domain typically predict higher accuracy for losses. In decision field theory (Busemeyer & Townsend, 1993), for example, choices in the loss domain are characterized by a slowing down of the evidence accumulation process dependent on the distance from the decision threshold, thus causing slower and more accurate responses. SSMs that assume a race between the evidence accumulation of competing options (e.g. S. D. Brown & Heathcote, 2008), also predict differences in accuracy. Finally, Hunt et al. (2012) proposed a biophysically plausible network model that predicted slower decisions when choosing between options with overall lower value. However, this model also did not account for the absence of accuracy effects.

A possible explanation of the increase in non-decision time in the loss domain and when less information is provided is that less advantageous contexts might provoke motor inhibition, similarly to a Pavlovian bias (Boureau & Dayan, 2011; Huys et al., 2011). This effect is also similar to the modulating function of the subthalamic nucleus in the basal ganglia circuit, which causes a "hold your horses" response (Frank, 2006b) in the presence of conflict. This would explain why responses could be delayed without affecting accuracy. Moreover, both effects of losses and partial feedback might not only be present in RTs, but also in meta-cognitive judgments like decision confidence. This idea is supported by a growing body of evidence showing how losses reduce confidence judgments in a variety of tasks (Lebreton, Langdon, et al., 2018; Lebreton, Bacily, Palminteri, & Engelmann, 2018).

A competing explanation might link the slowing down in the presence of losses to the loss attention framework (Yechiam & Hochman, 2013), i.e., the idea that losses receive more attention. However, increased attention has been previously linked to increases in the drift-rate and threshold parameters, and not in the non-decision time, since higher attention is typically accompanied by higher accuracy (Krajbich et al., 2010; Krajbich, Lu, Camerer, & Rangel, 2012). Moreover, Cavanagh et al. (2014) found that eye gaze dwell time only

predicted increases in the drift-rate towards the fixated option, independently on its value. They also found that pupil dilation was overall higher in the gain compared to the loss domain, and the relationship between pupil dilation and threshold was stronger in the gain compared to the loss domain.

In conclusion, RTs and accuracy are two behavioral manifestations of internal decision processes. These two variables provide complementary and equally important clues on the computations underpinning affective decision making, and should be jointly considered in order to build a comprehensive account of goal-directed behavior.

Chapter 3

A Reinforcement Learning Diffusion Decision Model for Value-Based Decisions

Laura Fontanesi, Sebastian Gluth, Mikhail S. Spektor, and Jörg Rieskamp

The manuscript in its current form has been accepted for publication in *Psychonomic Bulletin & Review*.

Financial disclosure: This research is supported by Grant 100014_153616/1 from the Swiss National Science Foundation.

Abstract: Psychological models of value-based decision making describe how subjective values are formed and mapped to single choices. Recently, additional efforts have been made to describe the temporal dynamics of these processes by adopting sequential sampling models from the perceptual decision making tradition, such as the diffusion decision model (DDM). These models, when applied to value-based decision making, allow to map subjective values not only to choices but also to response times. However, very few attempts have been made to adapt these models to situations in which decisions are followed by rewards, thereby producing learning effects. In this study, we propose a new combined reinforcement learning diffusion decision model (RLDDM) and test it on a learning task in which pairs of options differ with respect to both value difference and overall value. We found that participants became more accurate and faster with learning, responded faster and more accurately when options had more dissimilar values, and decided faster when confronted with more attractive (i.e., overall more valuable) pairs of options. We demonstrate that the suggested RLDDM can accommodate these effects and does so better than previously proposed models. To gain a better understanding of the model dynamics, we also compare it to standard DDMs and reinforcement learning models. Our work is a step forward towards bridging the gap between two traditions of decision-making research.

3.1 Introduction

Research on value-based decisions investigates how individuals value options and make decisions between them. Every-day decisions can be based on descriptive information, such as when choosing a restaurant based on reviews, or on personal experience, such as when choosing a restaurant based on previous visits. Reinforcement learning (RL, Sutton & Barto, 1998) describes the processes involved in the latter case, and specifically how the value associated with an option is updated following reward or punishment.

In the past decades, substantial progresses in understanding the mechanisms of RL have been made both in psychology (e.g., Estes, 1950; Luce, 1959; Bechara, Damasio, Damasio, & Anderson, 1994; Erev, 1998; Yechiam & Busemeyer, 2005; Rieskamp & Otto, 2006) and neuroscience (e.g., Schultz, Dayan, & Montague, 1997; Holroyd & Coles, 2002; Frank et al., 2004; Dayan & Daw, 2008; Niv, 2009). Within this framework, computational models can be used to infer latent value representations and psychological constructs (Lewandowsky & Simon, 2010), for instance, the reliance on more recent or past feedback (often referred to as the *learning rate*). RL models usually have two components: a learning component, that describes how past information is integrated with newly received feedback to update options' subjective values, and a choice model, that maps the subjective values associated with the options to the final choice probabilities. Despite providing a good fit to choice data, this mapping function (e.g., the soft-max choice rule) does not provide a description of the cognitive processes that lead to a specific decision. Fortunately, these mechanisms can be revealed by simultaneously inspecting choices and response times (RTs). For example, making the same choice faster or slower can indicate less or more decision conflict, respectively (Frank, Samanta, Moustafa, & Sherman, 2007). Furthermore, choices and RTs might be differently affected under different conditions, and those cognitive processes that only affect RTs would be overlooked by models based on choices alone.

Sequential sampling models (SSMs; for an overview, see Smith & Ratcliff, 2004; Bogacz et al., 2006) are process models that aim to describe the cognitive computations underlying decisions and allow predicting choices and RTs in a combined fashion. SSMs define decision making as an integration-to-bound process: When deciding between two options, noisy evidence in favor of one over the other option is integrated over time, and a response is initiated as soon as the evidence reaches a pre-set threshold. Cautious decision makers increase their threshold to make more accurate, but at the same time slower decisions. On the other hand, if the situation requires to respond as quickly as possible, the

threshold can be lowered at the cost of accuracy. When confronted with an easy decision (i.e., between a very good and a very bad option), the integration (or *drift*) rate is higher, leading to faster and more accurate decisions. SSMs have been successfully applied in many psychological domains (for an overview, see Ratcliff et al., 2016), including both perceptual and value-based decision making (e.g., Usher & McClelland, 2001; Busemeyer & Townsend, 1993). In particular, the diffusion decision model (DDM; Ratcliff, 1978), the dominant model in perceptual decision making, has gained particular popularity in value-based decision making research (Summerfield & Tsetsos, 2012). Thus, the DDM has been used to directly compare perceptual and value-based choices (Dutilh & Rieskamp, 2016), and it has been extended to account for and to model eye-movement data in consumer-choice behavior (Krajbich et al., 2010, 2012). Moreover, building on the discovery of a neural correlate of the integration-to-bound process during perceptual decisions in non-human primates (Gold & Shadlen, 2001), SSMs have also been used to link behavioral and neural measures, such as the decision threshold to activity in the striatum (Forstmann et al., 2008; Gluth et al., 2012; van Maanen et al., 2016).

While significant progress has been made in describing the processes underlying value-based decision making, previous work mainly focused on situations in which rewards are not provided after each choice. SSMs typically assume that the subjective value associated with some option is stable in time and that repeated choices involving the same option do not affect its subjective valuation. The assumption of stable preferences might hold in many choice situations, but is presumably violated when some kind of feedback is received after every choice. In these cases, SSMs should be extended by adding a learning component.

To overcome the limitations of both SSMs (i.e., the absence of learning processes) and RL models (i.e., the absence of mechanistic decision processes), new models need to be developed. The goal of the present work is to propose a new computational cognitive model that describes both the processes underlying a single decision (by relying on the SSM approach of decision making) and how these are influenced by the learning of subjective values of options over time (by relying on the RL framework). So far, only few attempts have been made to combine these two approaches (Frank et al., 2015; Pedersen et al., 2017). In particular, these studies have proposed variants of the DDM in which an RL rule is used to update the subjective values, and these values in turn are mapped to trial-specific DDM parameters in a meaningful way (e.g., the difference in subjective values is mapped to the drift rate). Notably, in these studies, only the reward differences between options were

manipulated, but not the mean values of different pairs of options. However, mean values have been reported to influence the speed of decisions (Polania et al., 2014; Palminteri et al., 2015; Pirrone, Azab, Hayden, Stafford, & Marshall, 2017), and could therefore be an important modulating factor of decisions during learning. Finally, an open question remains whether the subjective-value differences map linearly (as previously proposed) or non-linearly to the DDM parameters (more similarly to common decision rules in RL models, such as the soft-max choice rule).

In the present work, we propose a learning task in which not only value differences but also the mean values across different pairs of options are manipulated. We first test behavioral, cross-trial effects related to these manipulations, and develop a combined reinforcement learning diffusion decision model (RLDDM) that captures the observed learning and value-based behavioral effects. We then compare our model qualitatively and, whenever possible, quantitatively to other classes of models. We show that some of the value-based effects would have remained unnoticed if only choices but not RTs were taken into account—in particular those that are related to the mean value of pairs of options. Finally, we perform a rigorous model comparison analysis that illustrates the predictive advantages of the new model and provides insights into the cognitive processes underlying value-based decision making during learning.

3.2 Method

3.2.1 Participants and procedure

A total of 32 participants (24 female, age: 18-36, $M = 22.36$, $SD = 2.14$) completed the experiment. Participants were mainly psychology students recruited through the subject pool of the Faculty of Psychology of the University of Basel. Participation in the experiment was possible for partial fulfillment of course credits or cash (20 Swiss francs per hour). In addition, a monetary bonus corresponding to the performance in the experiment was awarded. Before starting the experiment, participants gave informed consent, as approved by the institutional review board of the Faculty of Psychology, University of Basel. The instructions of the task were presented directly on the screen. Information about participants' gender, age, handedness, and field of study were also requested on-screen before starting the task. Since an accuracy above 56% across 240 trials is unlikely due to random

behavior alone, according to a binomial test ($p < .05$), only participants who surpassed this threshold were included in the analyses. Raw data and scripts will be made available upon publication of the manuscript at `https://osf.io/95d4p/`.

3.2.2 Learning paradigm

The paradigm was a multi-armed bandit problem (Sutton & Barto, 1998). A total of four options per block were presented and participants chose between two of them in each trial. Options were randomly assigned either to the left or to the right of a fixation cross, and could be chosen by pressing either Q (for left) or P (for right) on the keyboard. After each choice, participants saw both options' rewards (i.e., full feedback) and collected the chosen option's reward. At the end of the experiment, the accumulated reward, divided by 1,400, was paid in Swiss Francs to the participants as a bonus (e.g., if they collected 7,000 points, they received 5 Swiss Francs). On average, participants gained a bonus of 8.10 Swiss francs.

Participants completed three experimental blocks of 80 trials, for a total of 240 trials. The payoffs of each option were not fixed but varied and were approximately normally distributed (Figure 3.1). The mean rewards of the options in each block were 36, 40, 50, and 54 for options A, B, C, and D, respectively. The standard deviation was 5 for all options. The payoffs were rounded to the unit, and were controlled to have representative observations (i.e., each participant observed the same outcomes in a different order, and the sample mean of each option was equal to the generating mean). The order of the payoffs of a single option was different in each block, and options were associated with four new visual stimuli (see below for a description of the visual stimuli), so that the options had to be learned again in a new block.

Each trial (Figure 3.2) was separated by a fixation cross, presented for 750–1,250 ms. The options were presented for up to 5,000 ms. If a response was faster than 150 ms or slower than 3,000 ms, the trial was discarded and a screen reminding to be slower or faster, respectively, was presented for 5,000 ms after the participant's response. Otherwise, the feedback was presented for 1,500 ms.

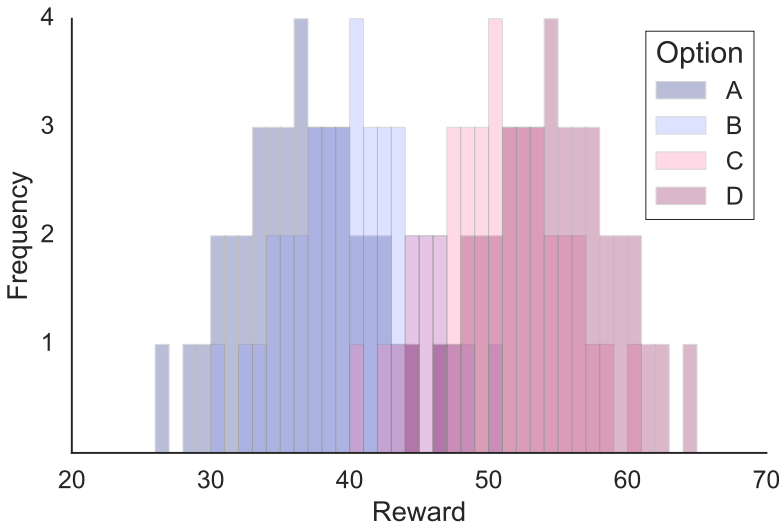


Figure 3.1: Reward distribution of the options A, B, C, and D in a learning block.

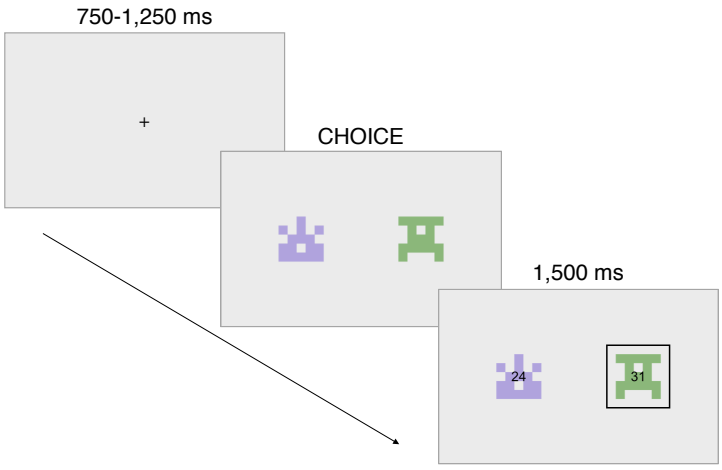


Figure 3.2: Example of a single trial: First, a fixation cross is shown from 750 to 1,250 ms; then, two of the four options are shown and a choice has to be made; finally, the reward corresponding to both options is presented, and the reward corresponding to the chosen option (highlighted by a black rectangle) is collected.

3.2.3 Design

In each learning block, only four of the six possible pairs of options were presented: AB, AC, BD, and CD (but not AD and BC). The order was pseudo-randomized so that the same pair would not be presented more than three times in a row. Presenting these four couples of options allowed us to test whether our model can predict two established behavioral effects of reward-based decision making in addition to the learning effects. Previous studies have shown that, when deciding among options that have similar values (i.e., difficult choices), people tend to be slower and less accurate (e.g., Polania et al., 2014; Dutilh & Rieskamp, 2016; Oud et al., 2016). We will refer to this effect as the *difficulty* effect. In our study, difficulty, given by the mean value difference, was low in pairs AC and BD (the difference was 14 on average), and high in pairs AB and CD (the difference was 4 on average). Previous studies have also shown that absolute shifts in value can affect decision speed without necessarily changing accuracy (e.g., Polania et al., 2014; Palminteri et al., 2015; Pirrone et al., 2017): Participants tend to be faster when deciding between two higher-valued options as compared to two lower-valued options. We will refer to this effect as the *magnitude* effect. In our study, magnitude, given by the mean value of the pairs of options, was lowest in pair AB (38), followed by AC (43), BD (47), and CD (52). Finally, we refer to the *learning* effect as the improvement in performance throughout the trials. In this study, each pair was presented for 20 trials per block, and each option was presented in 40 trials per block (since each option is included in two different pairs).

3.2.4 Stimuli

During the experiment, each participant saw a total of twelve different figures (four in each block) representing the options. The figures were matrices of 5×5 squares of which 17 were colored, arranged symmetrically along the vertical axis. To control for visual salience, we selected twelve evenly spaced colors in the $\mathrm{HSL_{UV}}$ space. A black rectangle was drawn around the chosen option at feedback presentation to highlight the collected reward. The experiment was programmed and presented using PsychoPy (Peirce, 2007).

3.2.5 Cognitive models

In total, we estimated three classes of computational models: RL models, the DDM, and combinations of the two, RLDDM (some of which were previously proposed by (Pedersen et al., 2017)). In the next sections, we present each class of models in detail.

3.2.5.1 Reinforcement learning models

RL models assume that the subjective values associated with the options are updated in each trial after experiencing a new reward (i.e., the reward feedback). These subjective values are then mapped to the probability of choosing one option over the other: Options with higher subjective values are chosen more often. Participants can differ in how much weight they give to new compared to old information: When more weight is given to old information, they are less affected by sudden changes in the rewards. They can also differ in how sensitive they are to subjective value differences: When they are very sensitive, their choices become more deterministic as they tend to always choose the option with the highest value. These two constructs, the stability of the subjective values and the deterministic nature of choices, are formalized in RL models by two parameters. The learning rate η (with $0 \leq \eta \leq 1$), and the sensitivity θ (with $\theta \geq 0$). The learning rate is the weight that is given to new information when updating the subjective value. When η is close to 0, the old subjective value remains almost unchanged (implying that even observations dating far back are taken into account), whereas when η is close to 1, the new subjective value almost coincides with the new information (implying that earlier observations are heavily discounted). The sensitivity parameter regulates how deterministic the choices are. With a higher θ, choices are more sensitive to value differences, meaning that subjectively higher-valued options will be chosen over lower-valued options with higher probability.

On each trial, the subjective values Q of the presented options are updated following the so-called *delta learning rule*:

$$Q_t = Q_{t-1} + \eta \cdot (f_t - Q_{t-1}) \tag{3.1}$$

where t is the trial number, and f is the experienced feedback. In the first learning block, Q-values were initialized at 27.5. This value was the average value shown in the task instructions at the beginning of the experiment, which was the same for all participants.

In the subsequent learning blocks, the Q-values were initialized at the mean values learned in the previous blocks. We reasoned that adjusting initial Q-values according to prior knowledge is more realistic than simply initializing them at zero. Indeed, preliminary model estimations revealed that all models provided better fits when adjusting Q-values to prior knowledge. Choices in each trial are predicted by the soft-max choice rule:

$$p_t = \frac{e^{\theta Q_{\text{cor}}}}{(e^{\theta Q_{\text{cor}}} + e^{\theta Q_{\text{inc}}})} \tag{3.2}$$

where p is the probability of choosing the option with the highest mean reward, and Q_{cor} and Q_{inc} are the subjective values of the options with a higher and lower mean reward, respectively.

Building on the simplest RL model, we took into account models that incorporate all possible combinations of two additional mechanisms, one concerning the learning rule and one concerning the choice rule. The first alternative mechanism allows η to differ depending on the sign of the reward prediction error. The reward prediction error is the difference between the feedback f_t and the previous reward expectation Q_{t-1}. Previous studies have found differences in learning rates for positive and negative reward prediction errors (Gershman, 2015) and have related this feature to optimism bias (Lefebvre, Lebreton, Meyniel, Bourgeois-Gironde, & Palminteri, 2017). The second mechanism allows θ to increase as a power function of how many times an option has been encountered before (as in Yechiam & Busemeyer, 2005) so that choices become more deterministic throughout a learning block:

$$\theta_t = (\frac{n}{b})^c \tag{3.3}$$

where n is the number of times an option has been presented, b (with $b > 0$) is a scaling parameter, and c (with $c \geq 0$) is the consistency parameter. When c is close to 0, θ reduces to 1 and is fixed in time, while higher values of c lead to steeper increase of sensitivity throughout learning.

3.2.5.2 Diffusion decision model

The DDM assumes that, when making a decision between two options, noisy evidence in favor of one over the other option is integrated over time until a pre-set threshold is reached. This threshold indicates how much of this relative evidence is enough to initiate a response. Since the incoming evidence is noisy, the integrated evidence becomes

more reliable as time passes. Therefore, higher thresholds lead to more accurate decisions. However, the cost of increasing the threshold is an increase of decision time. In addition, difficulty affects decisions: When confronted with an easier choice (e.g., between a very good and a very bad option), the integration process reaches the threshold faster, meaning that less time is needed to make a decision and that decisions are more accurate. The DDM also assumes that a portion of the RTs reflects processes that are unrelated to the decision time itself, such as motor processes, and that can differ across participants. Because of this dependency between noise in the information, accuracy, and speed of decisions, the DDM is able to simultaneously predict the probability of choosing one option over the other (i.e., accuracy) and the shape of the two RT distributions corresponding to the two choice options. Importantly, by fitting the standard DDM, we assume that repeated choices are independent of each other, and discard information about the order of the choices and the feedback after each choice. To formalize the described cognitive processes, the simple DDM (Ratcliff, 1978) has four core parameters: The drift rate v, which describes how fast the integration of evidence is, the threshold a (with $a > 0$), that is the amount of integrated evidence necessary to initiate a response, the starting-point bias, that is the evidence in favor of one option prior to evidence accumulation, and the non-decision time T_{er} (with $0 \leq T_{er} < \mathrm{RT}_{\mathrm{min}}$), the part of the response time that is not strictly related to the decision process (RT = decision time + T_{er}). Because, in our case, the position of the options was randomized to either the left or the right screen position, we assumed no starting-point bias and only considered drift rate, threshold, and non-decision time. Within a trial, evidence is accumulated according to the diffusion process, which is discretized in finite time steps according to:

$$x_{i+1} = x_i + \mathcal{N}(v \cdot dt, \sqrt{dt}), x_0 = a/2 \tag{3.4}$$

where i is the iteration within a trial, and a response is initiated as soon as $x \geq a$ or $x \leq 0$ (i.e., the evidence reaches the upper or the lower thresholds, respectively). The time unit dt is assumed to approach 0 in the limit (when $dt = 0$, the integration process is continuous in time). Choices are given by the value of x at the moment of the response (e.g., correct if $x \geq a$, incorrect if $x \leq 0$).

In total, we fit three versions of the DDM, varying in the number of free between-condition parameters. The first DDM had separate vs for difficult and easy choices, to allow accounting for the difficulty effect: Higher vs lead to faster and more accurate responses. The second model is as the first, but also has separate as for option pairs with a higher or lower mean reward. This model variant allows accounting for the magnitude effect: Lower

*a*s lead to faster, but not much more accurate decisions (Forstmann et al., 2011). This would explain the magnitude effect as a reduction of cautiousness: When confronted with more attractive options, individuals reduce their decision times (and therefore the time to the reward) by setting a lower threshold. The third model is as the second, but has also separate *v*s for option pairs with higher or lower mean reward, to check whether the magnitude effect is attributed only to a modulation of the threshold (i.e., cautiousness) or also to a modulation of the drift rate (i.e., individuals are better at discriminating two good options compared to two bad options).

3.2.5.3 Reinforcement learning diffusion decision models

The goal of our work is to propose a new model that overcomes the limitation of both the SSM and the RL frameworks. Therefore, we propose an RLDDM that is a combination of these two classes of models. The RLDDM simultaneously predicts choices and response times and describes how learning affects the decision process. Here, the DDM is tightly constrained by the assumed learning process: Instead of considering all choices as independent and interchangeable, the relationship between each choice, the experienced reward feedback, and the next choice is taken into account. The RLDDM assumes that, as in the RL framework, the subjective values associated with the options are updated after experiencing a reward feedback. The decision process itself is described by the DDM. In particular, the difference between the updated subjective values influences the speed of evidence integration in the next trial: When the difference is higher, as it might happen after experiencing several feedback, the integration becomes faster, leading to more accurate and faster responses. To formalize these concepts, we built a DDM in which the drift rate parameter is defined on each trial as the difference between the subjective values that are updated via the learning rule of RL models. The first and simplest RLDDM has four parameters (similarly to Model 1 in Pedersen et al. (2017)): one learning rate η to update the subjective values following Equation 3.1, a scaling parameter v_{mod} to scale the difference between values, one threshold a, and one non-decision time T_{er}. On each trial, the drift rate is defined as:

$$v_t = v_{\text{mod}} \cdot (Q_{\text{cor},t} - Q_{\text{inc},t}) \tag{3.5}$$

and within each trial evidence is accumulated as in Equation 3.4. Note that, since v is defined as the difference of subjective values, the difficulty effect naturally emerges from the model without assuming separate *v*s for easy and difficult choices.

We considered three additional mechanisms and fit different combinations of them, resulting in a total of eight different models. The first variation is similar to one considered for RL models and includes two separate ηs for positive and negative prediction errors (as in Pedersen et al. (2017)). The second variation is similar to one considered in the DDM to account for the magnitude effect. However, because subjective values are learned in time, instead of fitting separate as for different pairs of options (as we do in the DDM), we propose a trial-by-trial modulating mechanism:

$$a = exp(a_{\text{fix}} + a_{\text{mod}} \cdot \overline{Q}_{\text{pres}}) \tag{3.6}$$

where a_{fix} is the fixed threshold, a_{mod} is the threshold modulation parameter, and $\overline{Q}_{\text{pres}}$ is the average subjective value of the presented options. When $a_{\text{mod}} = 0$, this model reduces to the simplest model. The third variation is to make the mapping between subjective values and choices in the RLDDM more similar to the mapping in the soft-max choice rule. In Equation 3.5, v is linearly related to the difference in values. Since different pairs of options can have very similar or very different values (e.g., in Figure 3.1, pairs AB and AC), participants might differ in how sensitive they are to these differences. In RL models, this is regulated by the sensitivity parameter θ. We therefore propose a very similar, nonlinear transformation of the value differences in the definition of v:

$$v_t = S\big(v_{\text{mod}} \cdot (Q_{\text{cor},t} - Q_{\text{inc},t})\big), \tag{3.7}$$

with

$$S(z) = \frac{2 \cdot v_{\text{max}}}{1 + e^{-z}} - v_{\text{max}} \tag{3.8}$$

where $S(z)$ is an S-shaped function centered at 0, and v_{max} is the maximum absolute value that $S(z)$ can take on: $\lim_{z \to \pm\infty} S(z) = \pm v_{\text{max}}$. While v_{max} only affects the maximum and minimum values that the drift rate can take, v_{mod} affects the curvature of the function. Smaller values of v_{mod} lead to more linear mapping between the value difference and the drift rate, and therefore less sensitivity to value differences. Note that this model only resembles the previous models in the limit (i.e., when v_{max} has higher values).

3.2.6 Analysis of the behavioral effects

To assess the difficulty and the magnitude effects, we fit two separate Bayesian hierarchical models: a logistic regression on accuracy and a linear regression on log-transformed RTs. Accuracy was coded as 0 if the option with the lower mean reward was chosen (e.g.,

A is chosen over B), and as 1 if the option with higher mean reward was chosen (e.g., B is chosen over A). For both models, we included magnitude and difficulty as predictors and tested main effects and the interaction. Magnitude was defined as the true mean reward in each pair of options, and was standardized before fitting. Easy trials were coded as 1 and difficult trials as -1. For simplicity, and because we were interested in cross-trial effects, even though we were dealing with time-series data, no information about trial number was included in the models.

All models were fit using PyStan 2.18, a Python interface to Stan (Carpenter et al., 2017). We ran four parallel chains for 8,000 iterations each. The first halves of each chain were warm-up samples and were discarded. To assess convergence, we computed the Gelman-Rubin convergence diagnostic $\hat{R}$ (Gelman & Rubin, 1992). As an $\hat{R}$ close to 1 indicates convergence, we considered a model successfully converged when $\hat{R} \leq 1.01$. Weakly informative priors were chosen for both models. For a graphical representation of the Bayesian hierarchical models and for the exact prior distributions, see Appendix B.1.

To assess whether difficulty and magnitude had an effect on the behavioral data, we calculated the 95% Bayesian credible interval (BCI) on the posterior mean group distribution of the regression coefficients. If the BCI included 0, we concluded that there was no effect of a manipulation on either RT or choices. Finally, to assess model fit, we computed posterior predictive checks (Gelman et al., 1996) for mean accuracy and mean RT for each pair of options and looked whether the 95% BCIs of the posterior predictive distributions included the observed mean accuracies and RTs for AB, AC, BD, and CD. Posterior predictive distributions are useful to assess the quality of the models in their ability to predict patterns observed in the data. To approximate the posterior predictive distributions, we drew 500 samples from the posterior distribution, generated 500 independent datasets, and then computed the mean accuracy and mean RTs in each dataset, separately for choice pairs.

3.2.7 Model fitting and model comparison

For all classes of cognitive models, parameters were estimated using a Bayesian hierarchical modeling approach. Again, all models were fit using PyStan. Since the models vary in their complexity, the sampler was run for a different number of iterations. We first started with few samples (i.e., 1,000) and checked for convergence, reflected in $\hat{R} \leq 1.01$. If

the model did not converge, more samples were collected. We also checked for saturation of the maximum tree depth (considered satisfactory if less than .1%), energy Bayesian Fraction of Missing Information, and for divergences (considered satisfactory if less than .1%). Four parallel chains were run for all models and only the second half of each chain was kept for later analyses.

To assess the predictive accuracy of the models, we computed the widely applicable information criterion (WAIC, Watanabe, 2013). To compute the WAIC, we used the variance of individual terms in the log predictive density summed over the data points to correct for model complexity, as it approximates best the results of leave-one-out cross-validation (Gelman, Carlin, Stern, & Rubin, 2014). We also computed the standard error of the difference in the predictive accuracy of the best RLDDM, DDM, and among the models of Pedersen et al. (2017), using the R package loo (Vehtari, Gelman, & Gabry, 2017). This measure provides a better understanding of the uncertainty around the difference in WAIC scores. We then proceeded with the posterior predictive checks: Posterior predictives were calculated for mean accuracy and mean RT across learning by binning the trials within the learning blocks in eight groups of ten trials and across the pairs of options AB, AC, BD, and CD. As for the regression analyses, we sampled 500 parameter sets from the joint posterior distribution and generated 500 independent full datasets using those parameters. We then computed the mean accuracy and RTs in each dataset, separately for choice pairs and trial bins.

For a graphical representation of the Bayesian hierarchical models, and details about the prior distributions, see Appendix B.2. It has been shown that RL models can suffer from poor identifiability due to low information content in the data (Spektor & Kellen, 2018). To alleviate this concern, we conducted a parameter recovery study whose results can be found in Appendix B.4.

3.3 Results

Five participants were excluded for not reaching the minimum criterion of accuracy (see Method section), so that the data of 27 participants were included in the following analyses. The mean accuracy ranged from .43 to .53 ($M = .49$, $SD = .04$) for the excluded participants, and from .62 to .94 ($M = .81$, $SD = .08$) for the remaining ones.

3.3.1 Behavioral results

On average, participants showed substantial learning effects (Figure 3.3a and 3.3b): The higher-valued option was chosen more often throughout the trials (from $M = .71$ in the first 20 trials, to $M = .86$ in the last 20 trials), while at the same time responses became faster (from $M = 1.51$ s in the first 20 trials, to $M = 1.36$ s in the last 20 trials). They also showed difficulty and magnitude effects (Figure 3.3c and 3.3d): They were more accurate in easier compared to difficult choices ($M = .89$ compared to $M = .74$), while at the same time being faster ($M = 1.38$ s compared to $M = 1.46$ s); they were not more accurate in higher valued choice pairs compared to lower valued ones ($M = .81$ compared to $M = .81$), but they were faster ($M = 1.35$ s compared to $M = 1.48$ s).

To test difficulty and magnitude effects on accuracy and RTs across trials, we fit two regression models. Results from the logistic regression model on accuracy suggest that only difficulty, but not magnitude, had an effect on accuracy. There was no interaction between difficulty and magnitude on accuracy. In particular, participants were less accurate when choosing between AB and CD compared to AC and BD. The 95% BCI was higher than 0 (0.39 to 0.71, $M = 0.56$) for the mean group difficulty coefficient (meaning that easier decisions were more accurate), but it was around 0 for the magnitude coefficient (-0.27 to 0.16, $M = -0.05$) and for the interaction coefficient (-0.26 to 0.15, $M = -0.05$). To check whether the regression model predictions fit the data well, we used posterior predictive checks. In particular, we checked whether the regression model correctly predicts the mean accuracy across different pairs of options. As can be seen in Figure 3.4a, the regression model correctly predicts the observed pattern.

Results from the linear regression model on RTs suggest that both magnitude and difficulty as well as their interaction had an effect on RTs. In particular, participants responded faster when BD and CD were presented, compared to AB and AC. They were also faster in easy trials (pairs AC and BD) compared to difficult trials (pairs AB and CD) and this effect was stronger for less attractive options. The 95% BCI was lower than 0 for the group-level magnitude coefficient (-0.12 to -0.07, $M = -0.10$), for the difficulty coefficient (-0.04 to -0.02, $M = -0.03$) and for the interaction coefficient (-0.06 to -0.02, $M = -0.04$). Similarly to the previous regression model, we also checked whether the regression model correctly predicts the mean RTs across the different pairs of options. As can be seen in Figure 3.4b, the regression model correctly predicts the observed pattern.

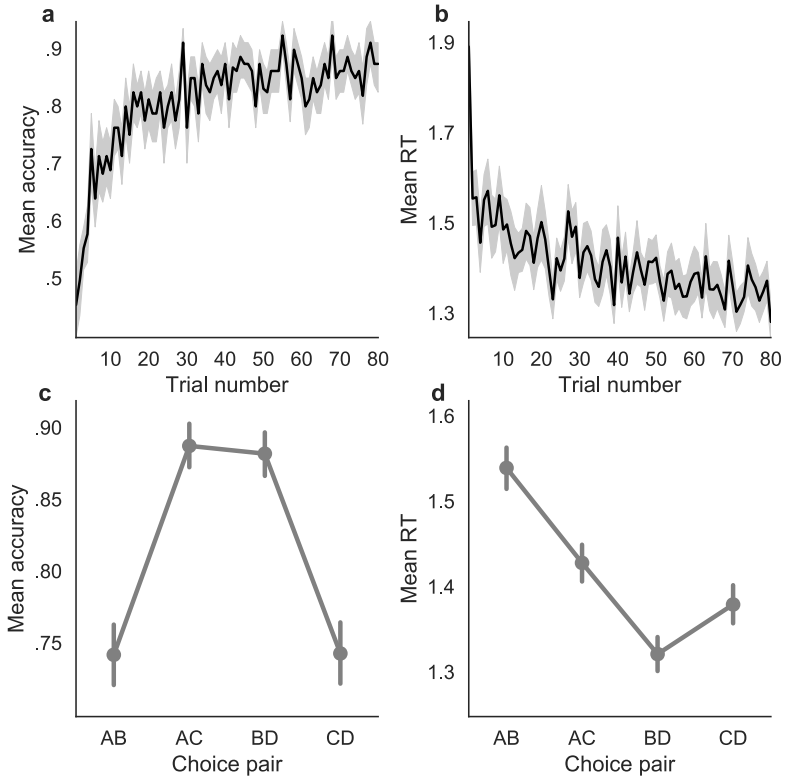


Figure 3.3: Mean accuracy (a) and RT (b) across participants as it develops throughout learning. Solid lines represent the mean across experimental blocks and participants, while the shaded areas represent 95% confidence intervals. Mean accuracy (c) and RT (d) across participants and for different pairs of options. Choices between options AC and BD were easier than between AB and CD, while the mean reward was highest in pair CD followed by BD, AC, and AB. The dots represent the mean across trials, while the bars represent 95% confidence intervals. Multilevel bootstrap was performed to account for repeated measures and therefore individual variability.

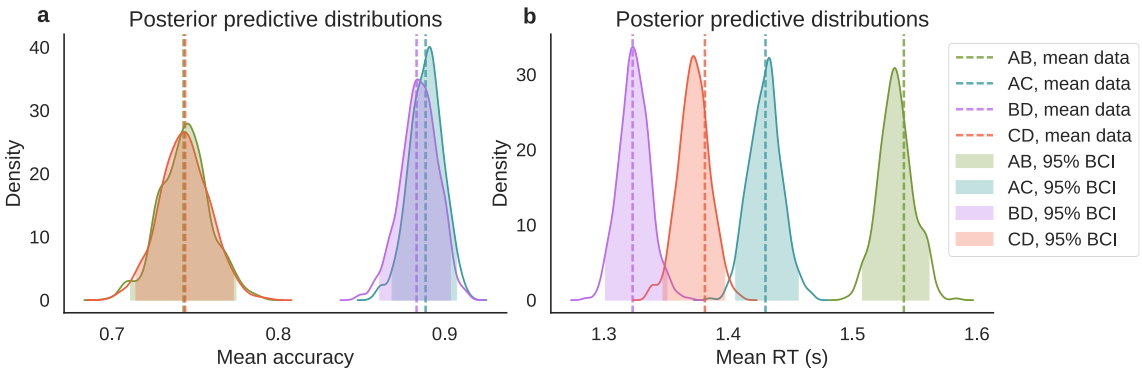


Figure 3.4: Posterior predictive distributions of mean accuracy (a) and mean RT (b) for different option pairs according to the logistic and linear regression models. The mean data (dotted lines) are compared to the regression model predictions (solid lines). The shaded areas represent the 95% Bayesian credible interval (BCI) of the posterior predictive distributions. Pairs AB and CD have similar mean values while pairs AC and BD have different mean value. Mean values increase from options AB, to AC, BD and CD.

3.3.2 Cognitive modeling

To better understand the learning and decision processes underlying value-based decisions, we fit and compared three different classes of models: RL models, the DDM, and RLDDMs, as well as previous attempts of combining RL and the DDM (Pedersen et al., 2017). While RL models can be only fit to choices, the DDM and RLDDM can be simultaneously fit to choices and RTs. However, the DDM does not take trial-by-trial information, such as the reward feedback, into account. In the following section, we report results from the model fitting and model comparison of these models.

Among the RL models, model 3 provided the most parsimonious account of the data. This model assumes separate learning rate parameters for positive and negative prediction errors and has a fixed sensitivity parameter throughout learning. Compared to the other models (Table 3.1), this model had the best predictive accuracy, as indicated by a higher log pointwise predictive density (lppd), and had lower complexity compared to the full model (i.e., model 4), as indicated by the p_{WAIC}. Judging by the WAIC, models 2 and 4, having an increasing sensitivity in time, did not outperform models 1 and 3, while the separate learning rates increased fitness of the models. This can be further assessed by looking at the 95% BCI of the posterior predictive distribution of mean accuracy across learning and pairs of options (Figure 3.5). All models predicted a nonlinear increase in performance throughout the trials, and a difference between easy (i.e., AC and BD) and difficult (i.e., AB and CD) choices.

The most parsimonious DDM was model 2, with separate drift rates for easy (i.e., AC and BD) and difficult (i.e., AB and CD) trials and separate response thresholds for pairs of options with different mean reward distributions (i.e., one for each pair: AB, AC, BD, and CD). As shown in Table 3.2, this model had lower predictive accuracy than model 3, as indicated by the lppd, but had also lower complexity than model 3, as indicated by the p_{WAIC}. The WAIC was lower for model 2 than for model 3, indicating that model 3 could not compensate its higher complexity with a better fit. Checking the posterior predictives in Figure 3.6, we can see that: (a) all three versions of the DDM did not predict any learning effect (i.e., both accuracy and RT are stable across trials); (b) only the versions of the DDM with separate thresholds for different option pairs could predict the magnitude effect on RTs, without changing accuracy predictions (i.e., having lower thresholds, responses in higher-valued pairs are not less accurate but only faster); (c) all models could predict difficulty effects on accuracy and RTs; (d) predictions from model 3 were not qualitatively

Table 3.1: Widely applicable information criteria of the reinforcement learning models.

Model	η	θ	p_{WAIC}	-lppd	WAIC
RL 1	one	fixed	48	2,636	5,368
RL 2	one	power	45	2,645	5,381
RL 3	two	fixed	63	2,569	5,265
RL4	two	power	72	2,573	5,291

Note. Models 1 to 4 are reinforcement learning (RL) models with learning rate η and sensitivity θ. The models could have a single or separate η (for positive and negative prediction errors). θ could be fixed in time or increase as a power function of the number of times an option was seen. p_{WAIC} is the effective number of parameters, lppd is the log predictive accuracy of the fitted model to data, and WAIC is the information criterion. Lower WAICs indicate better fits to data after accounting for model complexity.

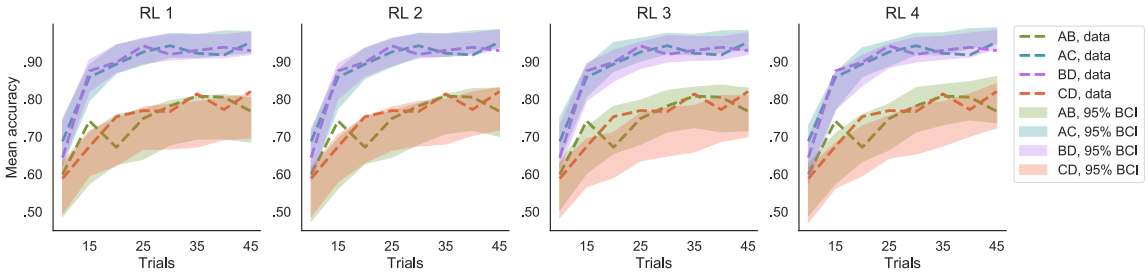


Figure 3.5: Posterior predictive distributions of mean accuracy according to the reinforcement learning (RL) models. The data (dotted lines) are compared to the 95% Bayesian credible interval (BCI) of the posterior predictive distribution (shaded areas), separately for the different options pairs and for 8 bins of trials within the learning blocks. Model 1 is the simplest RL model with one learning rate η and one sensitivity parameter θ. Models 3 and 4 have separate η for positive and negative prediction errors. In models 2 and 4, θ increases as a power function of the number of times an option is seen. According to the WAIC the best model is model 3.

Table 3.2: Widely applicable information criteria of the diffusion diffusion models.

Model	v	a	p_{WAIC}	-lppd	WAIC
DDM 1	difficulty	one	125	5,194	10,639
DDM 2	difficulty	all choice pairs	183	5,034	10,433
DDM 3	all choice pairs	all choice pairs	232	4,986	10,435

Note. Models 1 to 3 are diffusion decision models (DDMs) with drift rate v, decision-threshold a, and non-decision time T_{er}. v could depend either on choice difficulty only or different v could be fitted for each choice pair. a could be either fixed across conditions, or separate a could be fit for separate pairs of options.

better than predictions from model 2.

Among our proposed RLDDMs, the most parsimonious model was the last, full model. In this model, separate learning rates were fit for positive and negative prediction errors, the drift rate was an S-shaped function of the difference in subjective values, and the threshold was modulated by the learned average subjective value of the presented options, so that the threshold was lower when the expected reward was higher. As shown in Table 3.3, this model had highest predictive accuracy, as indicated by the lppd, and highest complexity, as indicated by the p_{WAIC}. Having the lowest WAIC suggests that the model's complexity is compensated by its superior fit to the data. In Figure 3.7, posterior predictives reveal how the different models were able to capture the observed patterns in accuracy and RTs. In particular: (a) all models were able to capture learning effects as a decrease in RTs and increase in accuracy over time; (b) all models captured difficulty effects, but only models 5 to 8, by including a non-linear mapping between value differences and drift rate, did not underestimate accuracy for difficult (i.e., AB and CD) decisions; (c) only the models that included a modulating effect of values on the decision threshold could capture the magnitude effect on RTs. While no significant qualitative pattern could be observed for two compared to one learning-rate models, all models with two learning rates had slightly lower WAICs compared to their analogues with only one learning rate. The best RLDDM also outperformed the best DDM, both in terms of WAIC and in terms of posterior predictive checks.

Among Pedersen et al. (2017)'s models, the most parsimonious one was a model with separate learning rates for positive and negative reward prediction errors, a drift rate

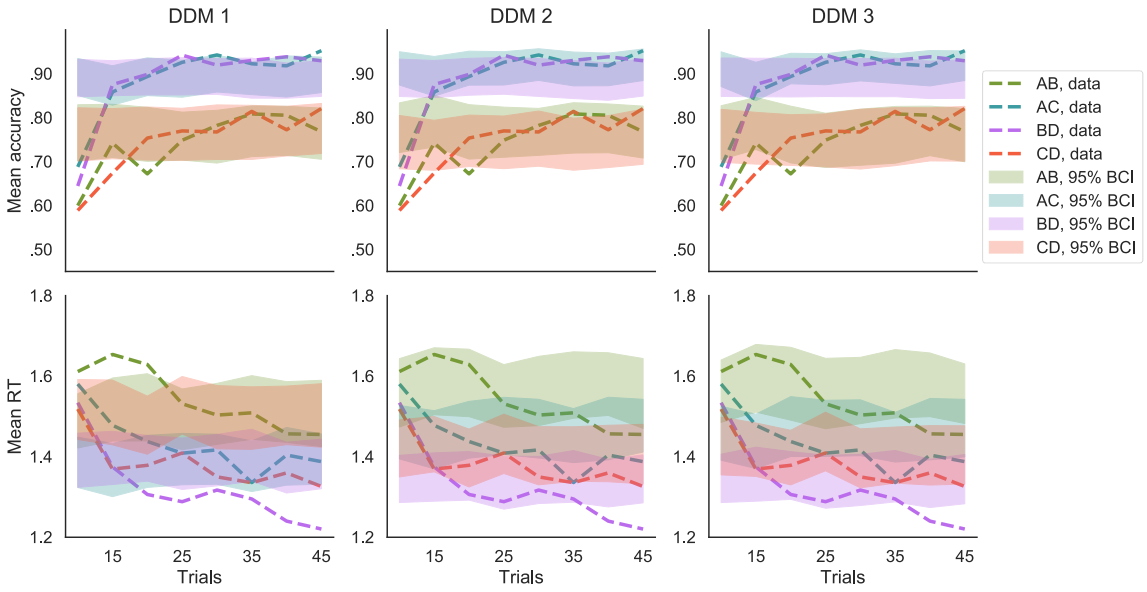


Figure 3.6: Posterior predictive distributions of mean accuracy and response time (RT) according to the diffusion decision model (DDM). The data (dotted lines) are compared to the 95% Bayesian credible interval (BCI) of the posterior predictive distribution (shaded areas), separately for the different options pairs and for 8 bins of trials within the learning blocks. Models 1 and 2 have separate drift rates v for easy and difficult decisions, while model 3 has separate v for each option pair. Model 1 has a fixed threshold a, while models 2 and 3 have separate a for each option pair. According to the WAIC the best model among DDM is model 2.

Table 3.3: Widely applicable information criteria of the reinforcement learning diffusion decision models.

Model	η	v	a	p_{WAIC}	-lppd	WAIC
RLDDM 1	one	linear	fixed	111	5,129	10,481
RLDDM 2	two	linear	fixed	134	5,051	10,369
RLDDM 3	one	linear	modulated	145	4,942	10,174
RLDDM 4	two	linear	modulated	159	4,866	10,048
RLDDM 5	one	sigmoid	fixed	137	4,930	10,135
RLDDM 6	two	sigmoid	fixed	159	4,861	10,039
RLDDM 7	one	sigmoid	modulated	164	4,672	9,672
RLDDM 8	two	sigmoid	modulated	190	4,613	9,607

Note. Models 1 to 8 are reinforcement learning diffusion decision models (RLDDMs) with learning rate η, decision threshold a, and non-decision time T_{er}. The models could have a single or separate η (for positive and negative prediction errors), linear or non-linear mapping of value differences to v, and fixed or value-modulated a.

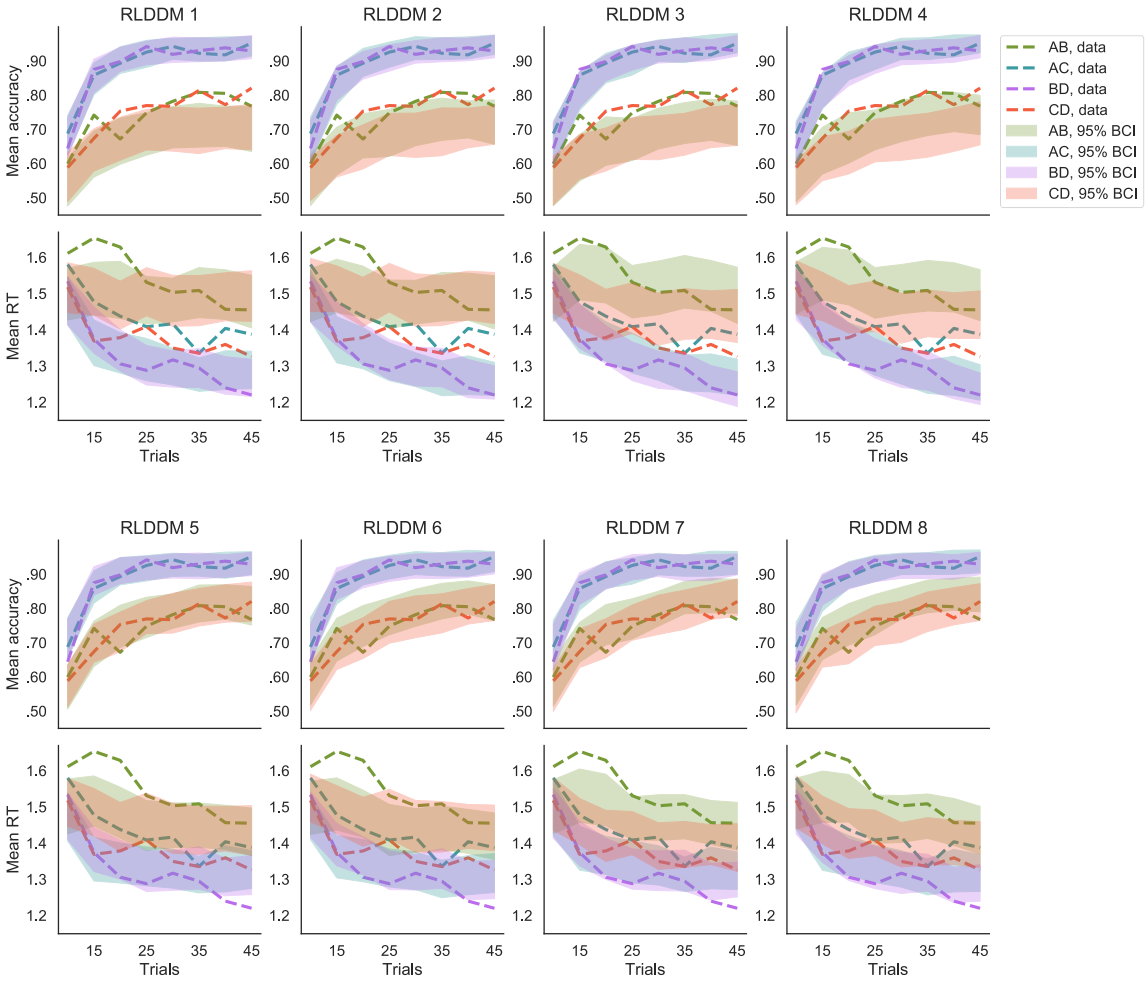


Figure 3.7: Posterior predictive distributions of mean accuracy and response time (RT) according to the reinforcement learning diffusion decision models (RLDDM). The data (dotted lines) are compared to the 95% Bayesian credible interval (BCI) of the posterior predictive distribution (shaded areas), separately for the different options pairs and for 8 bins of trials within the learning blocks. Models 1 to 4 have a linear mapping between differences in values and the drift rate v, and models 5 to 8 have a non-linear mapping. All models with even number have separate learning rate η for positive and negative prediction errors. Models 1, 2, 5, and 6 have a fixed threshold a, while, in models 3, 4, 7, and 8, a is modulated by the average value of the options. According to the WAIC the best model among DDM, RLDDM and the models of Pedersen et al. (2017), is model 8.

Table 3.4: Widely applicable information criteria of Pedersen et al. (2017)'s models.

Model	η	v	a	p_{WAIC}	-lppd	WAIC
Pedersen RLDDM 1	one	fixed	power	118	5,100	10,436
Pedersen RLDDM 2	one	power	power	112	5,326	10,875
Pedersen RLDDM 3	two	fixed	power	141	5,020	10,322
Pedersen RLDDM 4	two	power	power	126	5,240	10,732

Note. Models 1 to 4 are Pedersen et al. (2017) best fitting reinforcement learning diffusion decision models (RLDDMs) with learning rate η, decision threshold a, and non-decision time T_{er}. The models could have a single or separate η (for positive and negative prediction errors), fixed or increasing v, and fixed or decreasing a.

that is linearly proportional to the difference in values of the correct and incorrect options, and a threshold that decreases as a power function of time within a block. Note that this was also the most parsimonious model in their task. A quantitative comparison between the different combinations of models can be found in Table 3.4, while posterior predictives can be seen in Figure 3.8. The best of these models neither outperformed any of those RLDDMs that included the S-shaped mapping function in the drift rate, nor the ones having a modulating mechanism of value for the threshold.

Lastly, to have a measure of uncertainty of the difference in WAIC scores, we calculated the standard error of the difference in predictive accuracy of the best fitting RLDDM with the best fitting DDM, finding a substantial difference between the scores ($elpd_{\text{diff}} = -415.4$, $SE = 38.3$), and the best fitting model of Pedersen et al. (2017), finding a substantial difference between the scores ($elpd_{\text{diff}} = -255.1$, $SE = 33.1$).

3.4 Discussion

In the present work, we proposed a new process model for value-based decision making during learning. To test this model, we collected data from participants performing a multi-armed bandit task, in which both the value difference between options as well as the mean reward of different pairs of options were manipulated. This was done to elicit two value-based behavioral effects known in the literature: the difficulty (e.g., Polania et al., 2014; Dutilh & Rieskamp, 2016; Oud et al., 2016) and the magnitude (e.g., Polania et al.,

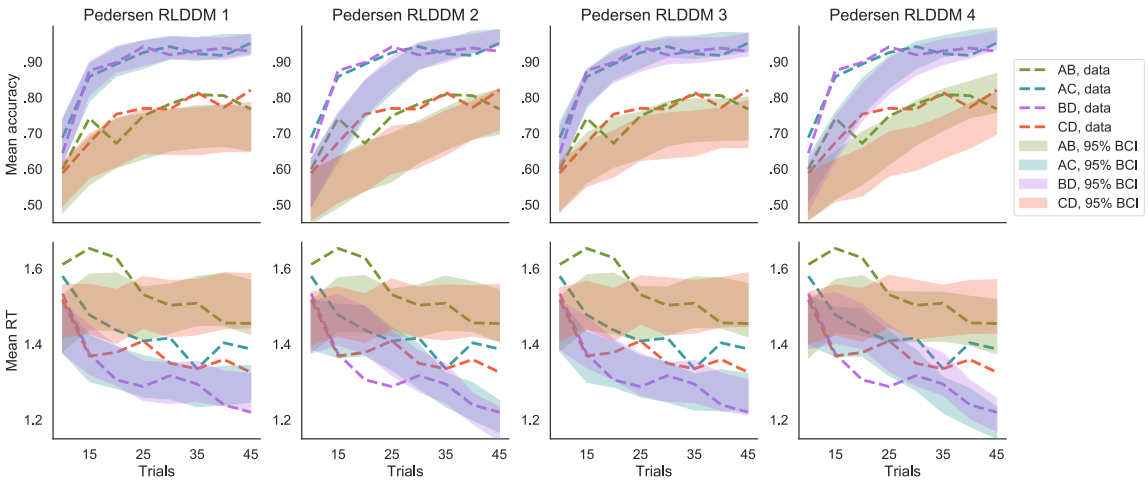


Figure 3.8: Posterior predictive distributions of mean accuracy and response time (RT) according to Pedersen et al. (2017) best-fitting models. The data (dotted lines) are compared to the 95% Bayesian credible interval (BCI) of the posterior predictive distribution (shaded areas), separately for the different options pairs and for 8 bins of trials within the learning blocks. Models 1 and 3 have a linear scaling parameter for the drift rate v, while in models 2 and 4 this parameter increases as a power function of the trial number. Models 3 and 4 have separate learning rate η for positive and negative prediction errors. In all models the threshold a decreases as a power function of the number of trials. According to the WAIC the best model among the models of Pedersen et al. (2017) is model 3.

2014; Palminteri et al., 2015; Pirrone et al., 2017) effects. We first assessed value effects across all trials by fitting regression models on accuracy and RTs. We observed a magnitude effect on RTs only and a difficulty effect on both RTs and choices. To gain insights into the separate learning and value mechanisms, we tested our model against RL models (from the value-based decision-making tradition) and standard DDMs (from the perceptual decision-making tradition). We also compared our model to a previously proposed class of combined RLDDM (Pedersen et al., 2017). Different classes of models were tested, when possible, quantitatively (i.e., whether our model provided a better account of the data using a relative measure) and qualitatively (i.e., whether our model captured the observed patterns that were related to the experimental manipulations).

Our analyses suggest that, while difficulty has an effect on both accuracy and RTs, magnitude only affects RTs: Difficult decisions tend to be slower and less accurate, while decisions among higher-valued options tend to be faster, but not less accurate. These results confirm previous studies that investigated value-based decisions after preferences have been formed (e.g., Polania et al., 2014; Pirrone et al., 2017) as well as studies that compared approach and avoidance behavior (e.g., Cavanagh et al., 2014; Palminteri et al., 2015). In line with previous studies, we also found that participants tended to become faster and more accurate during learning, and more so for easy compared to difficult trials. These behavioral patterns (a) can only partially be predicted by RL models, as they do not predict RTs, (b) are not predicted by the DDM, as it does not take trial-by-trial feedback into account, and (c) are fully predicted by RLDDM. By presenting easy and difficult pairs of options, we also showed that a nonlinear mapping between the difference in subjective values (learned via RL) and the DDM drift rate improved the model fit substantially. In other words, the drift rate may not double for option pairs whose difference of means is twice as large (for a similar finding in perceptual decisions, see Teodorescu, Moran, & Usher, 2015). As a consequence, models that do not assume a nonlinear mapping tend to underestimate the accuracy in the difficult trials. Finally, to give an account of the magnitude effect during learning, we proposed a mechanism in which the threshold is modulated by the mean subjective values of the presented options. By having a lower decision threshold, decisions between higher-valued pairs of options become faster, while accuracy only decreases to a minor extent. This mechanism is also suggested by the estimated thresholds in the DDM, separately fit for each pair of options: Higher-valued pairs have a lower threshold (see Figure 3.9a). In the RLDDM, this is obtained by a negative threshold modulation parameter: Negative values imply a lower threshold for higher-valued options (see Figure 3.9b). Cavanagh et al. (2014) also reported a reduction of the threshold when comparing approach to avoidance behavior,

and interpreted this finding as a facilitation effect on the cortico-striatal indirect pathway due to increased dopamine levels, based on previous work (Wiecki & Frank, 2013). In both RL and RLDDM models, separating the learning rate for positive and negative prediction errors increased the predictive accuracy of the models, as indicated by a lower WAIC. Although a qualitative difference in fit cannot be visualized in the posterior predictive checks we calculated, this result is in line with previous research (Gershman, 2015; Lefebvre et al., 2017).

Notably, our proposed model has stricter constraints than the DDM: When fitting the DDM to behavioral data, all trials are collapsed into two RT distributions for choosing high- and low-value options, meaning that slower decisions will be in the right tail of the distribution, independently of their occurrence at the beginning or at the end of a learning block. In the RLDDM, the trial order is taken into account, as the drift rate and the threshold depend on the learned values in each trial. When fitting the DDM using different parameters per condition, we also have less constraints. By explicitly relating the difference in values to the drift rate, and the mean learned values to the threshold, we propose mapping functions that can accommodate the observed results, providing more mechanistic explanations.

We also compared our best RLDDM to previously proposed RLDDMs. We fit the four best models proposed by Pedersen et al. (2017), and compared them quantitatively (see Table 3.4) and qualitatively (see Figure 3.8) to our models. The quantitative comparison confirmed that our best model had better predictive accuracy than those previous models. The qualitative comparison shows that the models proposed by Pedersen et al. (2017), which assume a linear mapping between value differences and drift rate, largely overestimate the difference in performance between easy and difficult choices. Models that assumed an increasing scaling parameter for the drift rate (i.e., models 2 and 4 in Figure 3.8) predicted an almost linear increase in accuracy throughout time, while the data suggest a more asymptotic learning curve. Moreover, all models proposed by Pedersen et al. (2017) predicted that the RTs for difficult (i.e., AB and CD) decisions do not decrease in time as much as it was observed in the data. Because Pedersen et al. (2017) did not show the development of mean RT throughout learning, we cannot assess whether this discrepancy was also present in their data. Finally, since Pedersen et al. (2017) did not manipulate mean reward of pairs of options and did not include a mechanism to account for the magnitude effect, their model is unable to explain this effect.

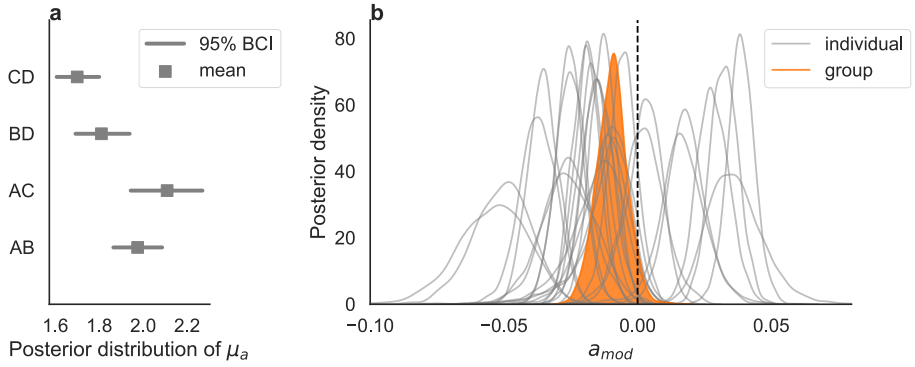


Figure 3.9: (a) Posterior distributions of the mean group threshold parameters μ_a of the diffusion decision model in which separate drift rates were fit for easy and difficult decisions, and separate thresholds were fit for all different pairs of choices. Solid lines represent the 95% Bayesian credible interval, and squares represent the mean of the posterior distribution. (b) Posterior distribution of the mean group (in orange) and of the individual (in grey) threshold modulator parameters a_{mod} of the full reinforcement learning diffusion decision model.

In this study, by combining DDM and RL models, we aimed at providing a psychological account of the processes underlying behavior in a RL paradigm, where the expected reward and the difficulty varies across trials. As behavioral effects in this task are not necessarily evident in all behavioral measures (in particular, the magnitude effect is only present in RTs), we showed how, by simultaneously fitting choices and RTs and constraining them by feedback data, we can provide a more complete account of the cognitive processes in this task and identify mechanisms that would remain undetected if only choices but not RTs were taken into account.

Future work should test whether the best RLDDM is also successful in describing behavior in different learning paradigms by, for instance, investigating the magnitude effect in the presence of gains and losses or by manipulating the dispersion of the reward distributions. Moreover, RLDDMs could be validated by linking model parameters to neural measures. Trial-by-trial variables such as prediction errors, reward expectation signals, trial-by-trial threshold modulations, and individual parameters such as learning rates for negative and positive prediction errors can be easily estimated using this model. It would be interesting to see whether the model predictions about prediction errors are in line with previous work in RL neuroscience (O'Doherty, Hampton, & Kim, 2007), and whether the trial-by-trial adjustments of decision thresholds are mapped onto the same brain circuitry which has been reported for decision-making paradigms without learning (van Maanen et al., 2011; Gluth et al., 2012; Gluth & Rieskamp, 2017). An obvious limitation of any RLDDM is that it can only be fit to two-alternative forced choice tasks. This is problematic for paradigms such as the Iowa gambling task (Bechara et al., 1994), for example, in which participants choose between four decks in each trial, or for studying context effects in experience-based decisions with more than two options (Spektor et al., in press). A different, multi-alternative SSM that is, for instance, based on the linear ballistic accumulator model (S. D. Brown & Heathcote, 2008) could offer an alternative.

In conclusion, by integrating knowledge from two separate research traditions, we showed how an extended computational model can promote our understanding of the cognitive processes underlying value-based decisions during learning.

Chapter 4

The Role of Dopaminergic Nuclei in Predicting and Experiencing Gains and Losses: A 7T Human fMRI Study

Laura Fontanesi, Sebastian Gluth, Birte Forstmann[1], and Jörg Rieskamp[1]

Financial disclosure: This research is supported by the Swiss National Science Foundation (mobility grant number P1BSP1_172017, and grant 100014_153616/1), the Department of Psychology of the University of Amsterdam, the Netherlands Organisation for Scientific Research (project number 14017), and an ERC grant from the European Research Council (B.U.F.).

[1] Shared senior authorship

Abstract: The ability to correctly predict the outcomes of actions based on previously experienced gains and losses is crucial for our success in a dynamic environment. Using invasive electrophysiological techniques in animal studies, it was found that the dopamine neurons, situated in the substantia nigra (SN) and the ventral tegmental area (VTA), have a crucial role in learning-by-feedback: They fire more or less depending on whether more or less rewards are delivered compared to previous expectations. Moreover, they modulate the activity of cortical and subcortical areas that are crucial for goal-directed-behavior. However, human studies using non-invasive neuroimaging methods, due to technical limitations, have almost exclusively investigated activity in target areas of dopamine neurons and provided inconclusive results. In this study, we used ultra-high field 7 Tesla magnetic resonance imaging (MRI) and optimized protocols to extract the signal of the SN and the VTA while human participants engaged in a gambling task. First, we found a significant overlap between individual VTA masks and previously proposed SN masks in MNI space. Therefore, the segmentation of these nuclei based on individual anatomy is crucial in order to obtain the anatomical precision for separating signals from adjacent deep-brain nuclei. Second, we found a significant correlation with the reward prediction error in both the SN and the VTA and no correlation with expected value, confirming the hypothesis that SN and VTA are responsible for learning-by-feedback. Moreover, activity in both the SN and the VTA correlated with risk: Their activity was higher when more certain outcomes were to be expected. Finally, we only found a surprise signal in the SN. This result is in line with a recent framework that proposed a differential role for the VTA and SN in learning, respectively, learning of values and learning of salience.

4.1 Introduction

In order to adapt to an ever changing environment, it is crucial to correctly predict the outcomes of our choices, as well as to update our expectations when they happen to be wrong. These learning processes were formalized within the reinforcement learning (RL) framework (Sutton & Barto, 1998), unifying the fields of psychology and artificial intelligence. In this framework, the reward prediction error (RPE) is defined as the difference between the expectations and the experienced rewards or punishments, and guides learning: New expectations are a weighted sum of past expectations and the RPE. By presenting participants in the lab with different options and providing feedback after every decision, psychologists and neuroscientists have investigated two main classes of cognitive processes. The first class consists of processes related to expectations. These are the expected value (EV), which is the mean expected outcome, and risk, which is the expected variance of the outcomes. The second class consists of processes related to the processing of the feedback. These are the deviation from previous expectations, or RPE, or the salience of the outcome, or surprise.

A highly distributed network related to expectations and feedback-processing was found in both the animal and the human brain. Electrophysiological studies in rodents and non-human primates showed that midbrain dopaminergic neurons (i.e., in the substantia nigra, SN – specifically in its pars compacta, SNc – and in the ventral tegmental area, VTA) fire more, equal, or less in association with a positive, zero, or negative RPE (Schultz, 1998, 2015), and their firing ramps up faster with increasing risk expectations (Fiorillo et al., 2003). Firing of cells in the SNc has also been associated with surprise (Matsumoto & Hikosaka, 2009). Because dopamine nuclei are more challenging to target using non-invasive neuroimaging techniques, studies using human participants mainly focused on dopamine target areas (Arias-Carrión, Stamelou, Murillo-Rodríguez, Menéndez-Gonzáles, & Pöppel, 2010). Neural correlates of the RPE have been found in the ventral striatum (VS) and an expected reward signal has been found in VS, amygdala, as well as in frontal areas such as the orbital frontal cortex (OFC) and the medial prefrontal cortex (MPFC) (for an overview see, e.g., O'Doherty, 2004; O'Doherty & Bossaerts, 2008; Clithero & Rangel, 2014; Bartra, McGuire, & Kable, 2013). Both VS and anterior insula (AI) were found to signal predicted risk and surprise (Preuschoff et al., 2006; Singer, Critchley, & Preuschoff, 2009; Fouragnan, Retzler, & Philiastides, 2018).

The measurement of small dopaminergic nuclei signaling using functional MRI

(fMRI) is very challenging. One challenge pertains to the higher concentration of iron in the SN. This high concentration causes differences in the magnetic properties of the SN compared to, for example, cortical areas, and asks for customized structural and functional MRI scanning protocols (e.g., reduced echo times). Another problem is the physiological noise in the fMRI data due to the proximity of these areas to major arteries and cerebrospinal fluid. Finally, their limited volume and distance from the receive elements of the scanner, combined with anatomical variability and standard procedures such as spatial smoothing, lead to a high risk of mixing signals from neighboring nuclei (Eapen et al., 2011; de Hollander et al., 2015, 2017).

Because of these challenges, only very few studies have directly measured activation of small dopaminergic nuclei in human participants. An exception was the study of Zaghloul et al. (2009): Using microelectrode recordings during deep brain stimulation surgery in Parkinsons disease patients, they found SN activation in line with the RPE. However, a few studies used fMRI and reported contradicting evidence. D'Ardenne et al. (2008) found positive but not negative RPE in the VTA. Pauli et al. (2015) found only a positive RPE in the SNc, a negative RPE in the pars reticulata of the SN (SNr), as well as a negative expected value signal in the SNr. Zhang et al. (2017) found that, while the medial part of the SN encoded RPE, the lateral and ventral parts encoded surprise.

To the best of our knowledge, previous studies with human subjects (1) have not compared the VTA and the SN activation, (2) have not looked at all the variables related to expectations and feedback processing (i.e., they did not always include EV, risk, RPE, and surprise); (3) have not addressed the above-mentioned fMRI-specific challenges. In particular, previous studies have used high-field 3 Tesla (3T) MRI, spatial smoothing, and did not draw individual masks to delineate the VTA or the SN. At the same time, however, Zhang et al. (2017) used a large dataset (n = 485), thus improving power in their analyses.

Ultra-high-field (UHF) 7 Tesla (7T) MRI can help to increase signal-to-noise ratio (SNR) and BOLD contrast-to-noise ratio (CNR), leading to a more refined spatial resolution without loss of power or need for spatial smoothing (van der Zwaag, Schäfer, Marques, Turner, & Trampel, 2015). In this study, we used UHF-fMRI in combination with scanning protocols tailored to extract signals from subject-specific masks of the midbrain to overcome some of the previous limitations and clarify the findings of previous studies, especially regarding the function of the VTA and the SN. By adapting the paradigm proposed by (Preuschoff et al., 2006), we also investigated important variables such as risk and surprise,

as well as EV and RPE, thereby targeting processes of both expectation and feedback processing.

4.2 Method

4.2.1 Participants and procedure

Twenty-seven participants [8 male (mean age=24.7, SD=5.0, min=19, max=35), 19 female (mean age=24.4, SD=4.7, min=19, max=35)] took part in the experiment. The study was approved by the ethics committee of the University of Amsterdam. All participants completed two separate sessions, one to obtain multimodal, 0.7 mm isotropic structural data, and one to obtain 1.5 mm isotropic functional data while engaging in a gambling task. All participants were recruited from the University of Amsterdam subject pool, via flyers and posters at the Spinoza center for Neuroimaging and at the Academic Medical Center in Amsterdam, and via advertisements in the magazine of the Dutch Parkinson Society. All participants were required to be MRI compatible, between 18 and 40 years old, right-handed, without previous history of psychiatric conditions or neurological diseases, and to have normal or corrected-to-normal vision. Before taking part in the sessions they gave written consent, and, before the second session, they also received written instructions for the behavioral task. Before going in the MRI scanner, they all completed a training session in which they could try the experiment on a computer, and were given a written questionnaire to test their comprehension of the probability of winning and losing in each scenario of the behavioral task. All participants were given 20 euros for the second session, were endowed with 10 more euros, and could win or lose up to 7 euros (either added to or subtracted from the initial endowment) based on their performance in the task, as explained below.

4.2.2 Data acquisition

All images were acquired at a Philips Achieva 7T MRI scanner, situated at the Spinoza Centre for Neuroimaging in Amsterdam (Netherlands), using a Nova Medical 32-channel head array coil. During the first session, participants could choose whether to watch a movie or not. During the second session, the gambling task was presented using PsychoPy

(Peirce, 2007).

Structural MRI. T_1-weighted, T_2^*-weighted, and Quantitative Susceptibility Mapping (QSM, Langkammer et al., 2012) images were simultaneously obtained using a multi-echo magnetization-prepared rapid gradient echo (ME-MP2RAGE) sequence (Metere, Kober, Möller, & Schäfer, 2017; Caan et al., 2018). The sequence parameters were: $T_{I,1}$ = 670 ms, $T_{I,2}$ = 3675.4 ms, $T_{R,1}$ = 6.2 ms, $T_{R,2}$ = 31 ms, $T_{E,1}$ = 3 ms, $T_{E,2}$ = [3, 11.5, 19, 28.5 ms], $T_{R,MP2RAGE}$ = 6778 ms, flip angle$_1$: 4, flip angle$_2$: 4,bandwidth: 404.9 MHz, acceleration factor SENSE: 2, FOV = 205 x 205 x 164 mm^3, acquired voxel size: .7 x .7 x .7 mm^3, acquisition matrix: 292 x 290, reconstructed voxel size: .64 x .64 x .70 mm^3, turbo factor: 150 (resulting in 176 shots). The total acquisition time was 19.53 min.

Functional MRI. The functional MRI protocol was an adaptation of Protocol 3 as reported by de Hollander et al. (2017), originally designed for a 7T Siemens scanner located at the Max Planck Institute for Human Cognitive and Behavioral Sciences in Leipzig, Germany. This protocol was used to optimize the temporal signal-to-noise (tSNR) in iron-rich nuclei in the human midbrain. The present protocol consisted of 2 runs of 719 volumes with 30 slices. The acquisition time was 23.97 min per run. Other parameters were T_R = 2,000 ms, T_E = 17 ms, flip angle: 60, bandwidth: 2226.2 Hz, voxel size: 1.5 x 1.5 x 1.5 mm ^{3}, FOV = 192 x 192 x 49 mm^3, SENSE acceleration factor, P-reduction (AP): 3, matrix size: 128 x 128. To acquire images with such TE, TR, and voxel-size, the protocol did not employ Fat suppression, and, to increase SNR, the protocol did not employ Partial Fourier. After the first run, an EPI image with opposite phase coding direction as compared to the functional scan was acquired to help correcting for geometric distortions due to inhomogeneities in the B0 field using the TOPUP technique during preprocessing (see below).

4.2.3 Gambling task

The gambling task used in the present study is an adaptation of the task by (Preuschoff et al., 2006). In each trial (Figure 4.1 A), two numbers were sampled one after the other between 1 and 5 without replacement. At the beginning of each trial, before seeing both numbers, participants were asked to bet which of the two numbers will be higher: They could win 5 euros if they were correct, and lose 5 euros otherwise. Participants were also instructed that the sampling was (pseudo-) random and that their choice could not influence sampling. The texts *"Second number is HIGHER."* and *"Second number is LOWER."*

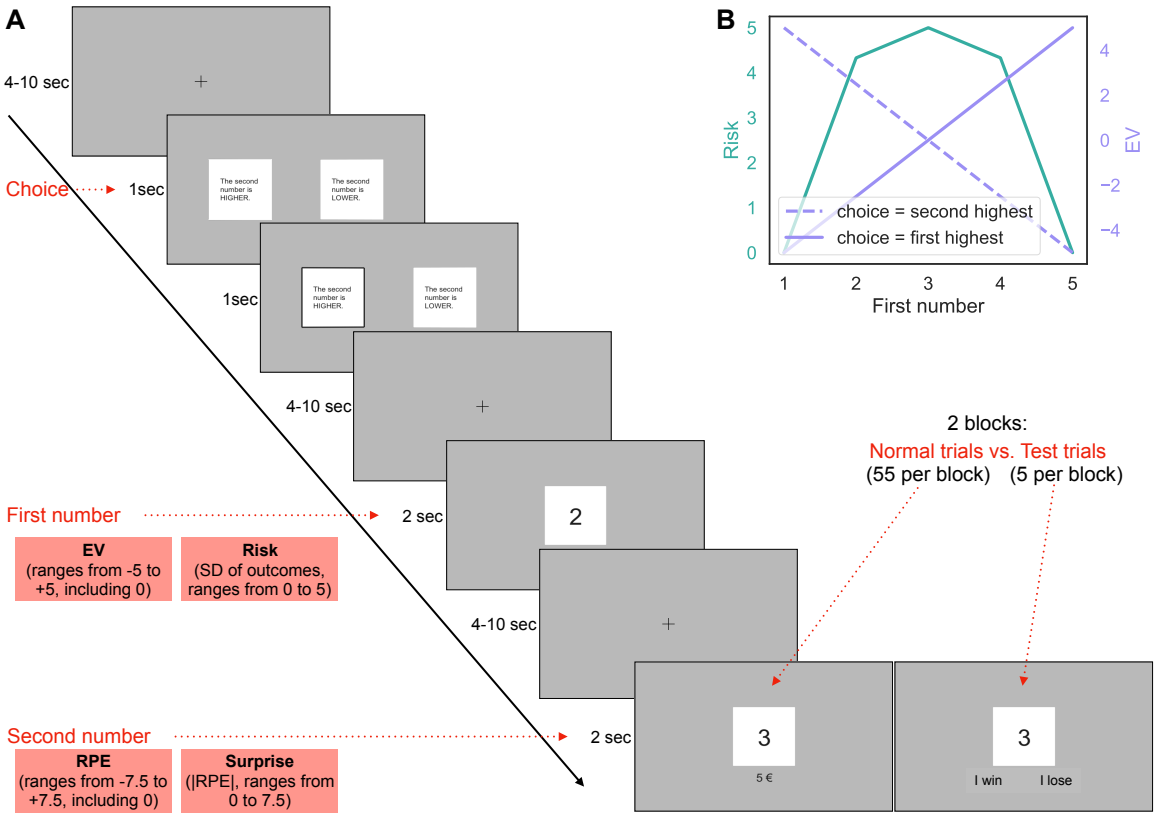


Figure 4.1: A. Example of a single trial. Between each event and at the beginning of each trial, a fixation cross is presented for a period of time between 4 and 10 seconds. A bet has to be placed within 1 second, and a rectangle is drawn around the corresponding choice for 1 more second. The first number is then shown for 2 seconds: In this example, the expected reward is 2.5 euros, and the risk is 4.3. Finally, the second number is shown for 2 seconds: In this case, both the reward prediction error and the surprise are 2.5. In test trials (8% of the total trials) participants have to specify whether they won or lost. B. Relationship between risk and expected reward when the first number is shown, depending on the choice.

appeared on the left and on the right sides of the screen (the position was counterbalanced across participants) and participants had to press either a left or a right button to place their bet. They could do so within 1 second, otherwise a bet would be placed for them at random. The choice (either the participant's or the random one) was then indicated by presenting a black frame around the corresponding text for another second.

The first number was subsequently shown for 2 seconds. At this point in time, the probabilities of winning and losing (both 50% at the beginning of the trial) change. For example, if a bet is placed on the second number being higher and the first number is 2, then three out of the four remaining numbers (i.e., 3, 4, and 5) lead to winning ($p_{winning} = 75\%$), while only one number (i.e., 1) leads to losing ($p_{losing} = 25\%$). The expected value of the gamble, is calculated as:

$$EV = p_{winning} \cdot 5 - p_{losing} \cdot 5 \tag{4.1}$$

and in this case is thus $5 \cdot 0.75 - 5 \cdot 0.25 = 2.5$ euros. The risk, defined as the variance of the possible outcomes (Markowitz, 1952), is thus 4.3. Note that, when the first number is 3, the probabilities of winning and losing remain 50%, the expected reward is always 0, and the risk is the highest, equal to 5. On the contrary, when the first number is either 1 or 5, participants already know whether they will lose or win (depending on what the bet was), therefore the expected value can be -5 or 5 euros and the risk is always 0. Since we were interested in neural correlates of both EV and risk, it is a crucial aspect of this design that EV and risk are not correlated (Figure 4.1 B).

At last, the second number is shown for 2 seconds, together with the corresponding gain or loss. At this point, the reward prediction error (RPE) is calculated:

$$RPE = outcome - EV. \tag{4.2}$$

In the example above (i.e., bet on 2nd card being higher; first card is 2), if the second number is 3, the reward is 5 euros and the reward prediction error is $5-2.5 = 2.5$ euros. The surprise, defined as the absolute value of the distance from the previous expectation (i.e., the reward expectation after the first card) as in Schultz (2015) and in Hayden, Heilbronner, Pearson, and Platt (2011), is thus $|5-2.5| = 2.5$. Since we were also interested in neural correlates of both RPE and surprise, it was also crucial that they were uncorrelated. This was the case, since RPE ranged between -7.5 and 7.5 and its distribution over trials was symmetrically centered around 0, and surprise was simply its absolute value.

The experiment consisted of 120 trials, divided in two blocks. In each block, 5 test trials were included to encourage participants to remain attentive throughout the experiment. In these trials, instead of showing the reward, we asked participants to indicate whether they won or lost. To correctly respond to this question, they needed to remember both their bet and the first number. At the end of the experiment, we randomly selected one of the 110 regular trials, and participants received the corresponding reward (i.e., 5 or -5 euros), plus 2 additional euros if they responded correctly to at least 8 of the 10 test trials, otherwise we subtracted 2 euros to the final reward. Between each event in each trial, and at the beginning of each trial, a fixation cross was presented for a period of time between 4 and 10 seconds, drawn from a truncated exponential distribution. The long inter-stimuli intervals were crucial to allow separating the BOLD signals associated with the first and the second numbers.

4.2.4 Behavioral analysis

Because choices are not influencing the chance of winning and losing in this task, behavioral analyses had the purpose to check the quality of the data for the fMRI analyses. The most important indicator of data quality was the accuracy in the test trials: Blocks in which participants made more than two out of five mistakes were discarded, where misses also counted as mistakes. Another important indicator was the number of missed bets: Blocks in which participants missed more than ten out of 60 bets were discarded. Finally, we checked the percentage of right vs. left responses. Because the position of the texts corresponding to the specific bets was counterbalanced across – but fixed within – participants, a similar number of right and left responses needed to be made for a balanced design. Blocks in which participants made less than ten right or more than fifty right (out of 60) choices were discarded.

4.2.5 Structural and functional MRI data preprocessing

Registration and preprocessing were performed using FMRIPREP version 1.0.6 (Esteban et al., 2018), a Nipype (Gorgolewski et al., 2011) based tool. Registration across session was done by registering the functional images (from the second session) to the T_1-weighted structural image multiplied by the first echo of the T_2^*-weighted structural images (from the first session). Because the T_1-weighted, T_2^*-weighted, and QSM structural images

were acquired simultaneously during the same scan in the first session, there was no need to co-register them first.

Structural images were corrected for intensity non-uniformity using N4 Bias Field Correction (Tustison et al., 2010) and skull-stripped using antsBrainExtraction.sh. Spatial normalization to the ICBM 152 Nonlinear Asymmetrical template (Fonov, Evans, McKinstry, Almli, & Collins, 2009) was performed through nonlinear registration with the antsRegistration tool of ANTs v2.1.0 (Avants, Epstein, Grossman, & Gee, 2008), using brain-extracted versions of both T_1-weighted volume and template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T_1-weighted image using *fast* (FSL v5.0.9) (Zhang, Larcher, Misic, & Dagher, 2001). Functional data was motion corrected using mcflirt (FSL v5.0.9, Jenkinson, Bannister, Brady, & Smith, 2002). Distortion correction was performed using an implementation of the TOPUP technique (Andersson, Skare, & Ashburner, 2003) using 3dQwarp (AFNI v16.2.07, Cox, 1996). This was followed by co-registration to the corresponding T_1-weighted image using boundary-based registration (Greve & Fischl, 2009) with 9 degrees of freedom, using flirt (FSL). Motion correcting transformations, field distortion correcting warp, BOLD-to-T_1-weighted transformation and T_1-weighted-to-template (MNI) warp were concatenated and applied in a single step using antsApplyTransforms (ANTs v2.1.0) using Lanczos interpolation.

Physiological noise regressors were extracted applying CompCor (Behzadi, Restom, Liau, & T.Liu, 2007). Principal components were estimated for the two CompCor variants: temporal (tCompCor) and anatomical (aCompCor). A mask to exclude signal with cortical origin was obtained by eroding the brain mask, ensuring it only contained subcortical structures. Six tCompCor components were then calculated including only the top 5% variable voxels within that subcortical mask. For aCompCor, six components were calculated within the intersection of the subcortical mask and the union of CSF and WM masks calculated in T_1-weighted space, after their projection to the native space of each functional run. Framewise displacement (FD, Power et al., 2014) was calculated for each functional run using the implementation of Nipype.

The preprocessing and registration output was visually inspected for each subject using the html output files of FMRIPREP. Functional data quality was assessed using MRIQC (Esteban et al., 2017) prior preprocessing, to check for visual artifacts and excessive head movements. Finally, after preprocessing and registration, tSNR maps were computed

using Nipype to assess the tSNR across the region of interests (ROIs).

4.2.6 Anatomical segmentation

One main aim of the present study was to obtain anatomically precise masks in the individual space for the two ROIs: the ventral tegmental area (VTA) and the substantia nigra (SN). Because of its relatively high iron concentration, the SN is most discernible in QSM images (Keuken et al., 2014), as shown in the first row of Figure 4.2. Unlike the SN, the VTA lacks clear anatomical borders. Segmentation can be performed, however, by exclusion from the neighboring iron-rich nuclei (i.e., the SN and the red nucleus, RN) and the CSF, so both should be clearly visible. The CSF is not visible in the QSM image. It is, however, clearly visible in the T_1-weighted image (see Figure 4.2, third row). To ease and improve the segmentation process, we therefore combined the T_2^*-weighted and T_1-weighted images, by first normalizing them within the midbrain area (i.e., a pre-selected area of 1.6 x 1.6 x 3.08 cm^3) and finally summing them up. The result can be seen in the bottom row of Figure 4.2.

Manual segmentation was performed using FSLView version 3.0.2, by two independent and trained researchers (one of which is the first author of the current manuscript). Only the voxels that were marked by both researchers were kept in the final masks, i.e., the conjunction masks. To assess inter-rater reliability (i.e., the agreement between the two researcher), we computed the Dice score (Dice, 1945) separately for each participant, hemisphere, and structure. The Dice score is computed as the ratio between the union of the two areas and the conjunction of the two areas. It therefore depends on the average dimension of the structure (with smaller structures having smaller scores) and has to be interpreted accordingly. Scores approaching 1 indicate good agreement between raters, while scores close to 0 indicate poor agreement between raters.

Drawing individual masks for each subject and area is a time- and resource-consuming process: High resolution structural images need to be acquired first, and then two trained researchers need to complete a lengthy segmentation process. To forgo this costly approach, SN (Keuken et al., 2014) and VTA (Pauli et al., 2018) MRI atlases have been published in recent years. These atlases consists of probabilistic maps of different ROIs in MNI space, and can be thus transformed in the individual space to extract the signal from these regions. The disadvantage of this less resource-intensive approach is a

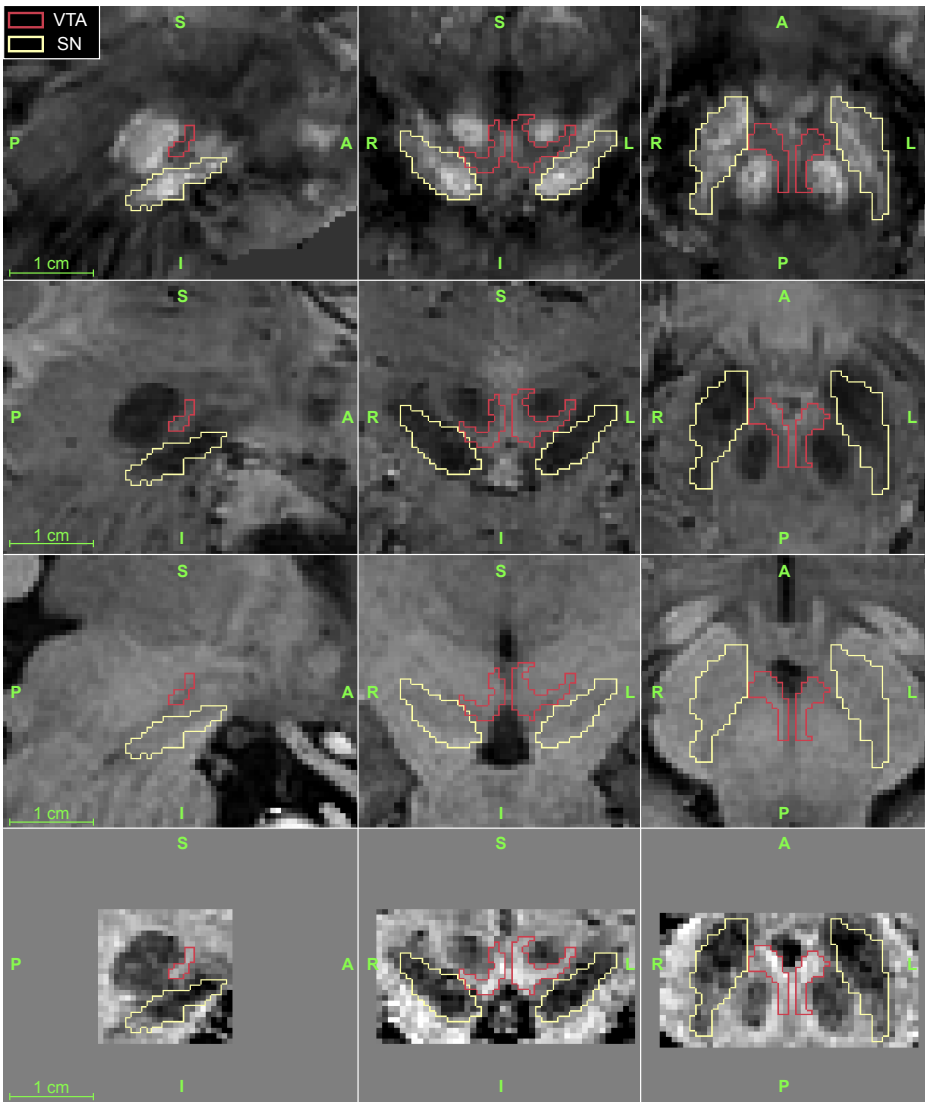


Figure 4.2: Anatomical images of the midbrain area of one participant in the sagittal (first column), coronal (second column), and axial (third column) planes. The first row is the QSM image, used for SN segmentation. The second and third row are, respectively, the average between the third and fourth echo of the T_2^*-weighted, and the T_1-weighted images. To obtain the image in fourth row, the images in the second and third row were normalized within the midbrain area (the non-homogeneous grey area in the last row) and then summed. This image was used for VTA segmentation, as it shows a contrast of both iron-rich nuclei and of the CSF.

potential loss of sensitivity and specificity due to disalignment between the individual and the standard spaces, as well as individual differences. To quantify the loss of information in this process, we transformed the three SN subregions proposed by Zhang et al. (2017), based on the 33% thresholded probabilistic masks proposed by Keuken et al. (2014), to the individual space and measured the overlap with our individual VTA masks as the number of voxels in common, divided by the overall area. A similar procedure was done with the proposed VTA and SN subdivisions of Pauli et al. (2018), using their deterministic atlas (50% thresholded).

4.2.7 fMRI data analysis

We extracted the fMRI signal for each time point within the ROIs (i.e., left and right SN and VTA) for each subject and computed its average time course for each ROI separately. We then fitted a GLM to the resulting time series for every region, participant, and block using statsmodels (Seabold & Perktold, 2010). Specifically, we used the GLSAR AR(1) model, to account for autocorrelation. The design matrices were constructed using Nistats (`https://nistats.github.io/index.html`). In the design matrices, the following events were convolved with the canonical, double-gamma hemodynamic response function (HRF), together with their temporal derivatives: the bet at the beginning of the trial, the appearance of the first card, the appearance of the second card in regular trials, and the appearance of the second card in test trials. On top of these, we added four parametric regressors: EV and risk (with onsets at the appearance of the first card and as amplitude the normalized EV and risk of each trial), and RPE and surprise (with onsets at the appearance of the second card and as amplitude the normalized RPE and surprise of each trial). Additional nuisance parameters were the six aCompCor, FD, six head movement variables provided by *fmriprep*, and cosine regressors for high-pass temporal filtering. For these analyses no spatial smoothing was used. After averaging across blocks, we performed independent two-sided t-tests, separately by ROIs and hemisphere (i.e., left vs. right) for the mean of the parameters corresponding to EV, risk, RPE, and surprise being equal to zero. We also estimated the equivalent Bayesian t-tests, as implemented in the BayesFactor R library (`https://cran.r-project.org/web/packages/BayesFactor/index.html`), as it quantifies evidence not only for the alternative, but also for the null hypothesis and therefore complements the frequentist analyses.

For the exploratory and control analyses, we estimated the same GLMs as on the

ROIs, using a mass-univariate, voxel-wise approach with Nistats (`https://nistats.github.io/index.html`). At the level of individual runs, we used a smoothing Gaussian kernel with a FWHM of 3.0 mm. At the participant level, we estimated the size of the baseline contrasts of the parameter estimates of EV, risk, RPE, and surprise. These participant-wise contrasts of parameter estimates (COPE) were then transformed to the MNI space and used in the third and final group-level analysis. Finally, we performed a Gaussian Random Field cluster analysis on the resulting four z-maps (EV, risk, RPE, and surprise), using FSL *cluster tool*. For these analyses, we set an input threshold of 2.3 and a cluster-wise threshold of $p<.05$.

4.3 Results

4.3.1 Quality of data assessment

Three blocks (from three different participants) were discarded based on behavior. One block was discarded because three out of the five test trials were incorrect, and the other two blocks were discarded because twelve out of sixty missed bets. In the remaining blocks, and over the two blocks, participants made on average 1.0 mistakes (SD=1.05, min=0, max=4), missed on average 4.48 trials (SD=3.65, min=0, max=12), and chose on average the right option on 57.81 trials (SD=13.75, min=21, max=88).

Two blocks (from two different participants) were discarded based on excessive head movements (mean FD over .3 mm). Because one of these blocks was already discarded based on behavior, a total of four blocks was excluded from the final analyses. In the remaining blocks, and over the two blocks, participants had an average mean FD of .14 mm (SD=.06, min=.04, max=.27).

The tSNR across ROIs can bee seen in Figure C.1

4.3.2 Anatomical masks

The Dice scores, measuring the inter-rater reliability, can be seen in Table 4.1. In general, higher scores were obtained for the SN as compared to the VTA. This is not surprising, considering that Dice scores are sensitive to overall size (the SN is approximately

3.7 times bigger than the VTA), and that the VTA lacks clear borders. By only keeping those voxels that both raters agreed on (i.e., the conjunction masks), we ensured that the voxels included in the analyses lie exclusively in the investigated ROI.

Table 4.1: Anatomical segmentation results.

			Mean	SD	Min	Max
SN	Right	Dice score	0.85	0.04	0.73	0.91
		Size (mm^3)	520.77	76.75	311.49	637.65
	Left	Dice score	0.84	0.04	0.74	0.90
		Size (mm^3)	501.67	60.86	384.26	621.25
VTA	Right	Dice score	0.56	0.07	0.43	0.68
		Size (mm^3)	138.91	39.37	76.51	233.26
	Left	Dice score	0.56	0.06	0.38	0.68
		Size (mm^3)	137.46	38.30	80.82	224.34

Note. Dice scores and size of the individual conjunction masks of the regions of interest (ROI): left and right substantia nigra (SN) and left and right ventral tegmental area (VTA). Conjunction masks are the intersection of the two independent raters' masks. Dice scores closer to 1 indicate higher agreement between the two raters, while dice scores close to 0 indicate lower agreement between the two raters.

In addition to the Dice scores, we also calculated the percentage of overlap between our individual-level conjunction masks and previously proposed group-level subdivisions of the SN and the VTA[2] (Zhang et al., 2017; Pauli et al., 2018), transformed to the individual space (see Figure 4.3). We found significant overlap between the medial parts of the SN and our individual VTA masks. Specifically, there was a mean overlap of 7.23 percent (SD=10.14, min=0.00, max=34.58, $p<0.001$) with the medial part of the SNc (mSNc), and a mean overlap of 1.3 percent (SD=2.14, min=0.00, max=8.36, $p<0.001$) with the lateral part of the SNc (lSNc) as defined by Zhang et al. (2017); and a mean overlap of 1.56 percent (SD=2.21, min=0.00, max=11.93, $p<0.001$) with the SNc as defined by Pauli et al. (2018). We also found a significant overlap between Pauli et al. (2018)'s subdivisions of the VTA (i.e., labelled VTA and PBP, where VTA is the more medial and PBP is the more lateral part) and our individual SN masks. Specifically, there was a mean overlap of 7.76 percent (SD=9.81, min=0.00, max=58.76, $p<0.001$) with Pauli et al. (2018)'s VTA and a

[2] Defined as the ratio between the number of voxels in common and the total number of voxels of the two regions.

mean overlap of 8.81 percent (SD=6.47, min=0.00, max=25.76, p<0.001) with Pauli et al. (2018)'s PBP.

Figure C.2, C.3, and C.4 show, respectively, a comparison between Pauli et al. (2018)'s atlas with our probabilistic VTA and SN maps in the MNI space, a comparison between Zhang et al. (2017)'s atlas with our probabilistic VTA and SN maps in the MNI space, and a comparison between Pauli et al. (2018)'s and Pauli et al. (2018)'s atlases in the individual space of one example subject.

4.3.3 ROI-wise GLM

Results of the ROI-wise GLM are shown in Table 4.2 and Figure 4.4. At the time of presentation of the first card, there were no parametric correlations between signal in any of the ROI with the EV, with the Bayes Factor (BF) pointing to substantial [3] evidence for the null hypothesis. However, there were significant correlations with risk in both the left-VTA (p<0.05) and the left-SN (p<0.05), with the BF pointing to strong evidence towards the alternative hypothesis. At the time of presentation of the second card, there were significant correlations with RPE in the left- and right-VTA (p<0.05), with the BF providing substantial support for the alternative hypothesis, and in the right-SN (p<0.05), with the BF providing weak support for the alternative hypothesis. Finally, we found a correlation with surprise in the right-SN (p<0.05), with the BF providing weak support for the alternative hypothesis, and no effect in the VTA, with the BF providing substantial support for the null hypothesis. Taken together, both VTA and SN were linked to risk before the outcome was revealed as well as to RPE after the outcome was revealed. Only SN was additionally associated with surprise about the outcome.

4.3.4 Voxel-wise GLM

Results of the voxel-wise GLM are shown in Table 4.3 and Figure 4.5. After cluster correction, we found positive correlations with EV in the orbitofrontal cortex and posterior cingulate cortex and negative correlations with EV in thalamus and anterior insula. We found positive correlation with risk in the superior temporal gyrus, anterior cingulate cortex

[3] see Jarosz and Wiley (2014)

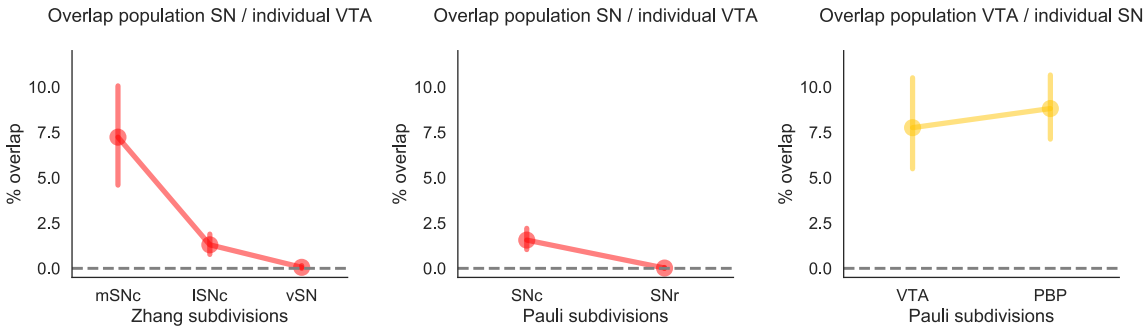


Figure 4.3: In red: percentage of overlap across our individual ventral tegmental area (VTA) masks (the conjunction mask across two independent raters) and substantia nigra (SN) subdivisions as defined by Zhang et al. (2017) and Pauli et al. (2018). The medial parts of the SN (mSNc, and SNc) overlap more with the VTA than the lateral and ventral parts of the SN (lSNc, vSN, and SNr). In yellow: percentage of overlap across our individual SN masks and VTA subdivisions as defined by Pauli et al. (2018). Both the ventral (VTA) and lateral (PBP) parts of the VTA overlap with the SN. Bars represent 95% confidence intervals.

Table 4.2: ROI-wise GLM results.

ROI	EV	risk	RPE	surprise
SN-left	$t(26)$=-0.16, p=0.87 BF_{10}=0.21	$t(26)$=-3.12, p=0.004* BF_{10}=9.48	$t(26)$=1.41, p=0.17 BF_{10}=0.50	$t(26)$=0.34, p=0.74 BF_{10}=0.21
SN-right	$t(26)$=0.27, p=0.79 BF_{10}=0.21	$t(26)$=-1.20, p=0.241 BF_{10}=0.39	$t(26)$=2.26, p=0.03* BF_{10}=1.76	$t(26)$=2.33, p=0.03* BF_{10}=2.01
VTA-left	$t(26)$=-0.49, p=0.63 BF_{10}=0.23	$t(26)$=-3.52, p=0.002* BF_{10}=22.19	$t(26)$=2.97, p=0.01* BF_{10}=6.81	$t(26)$=-0.32, p=0.75 BF_{10}=0.21
VTA-right	$t(26)$=0.07, p=0.94 BF_{10}=0.21	$t(26)$=-1.31, p=0.202 BF_{10}=0.44	$t(26)$=2.66, p=0.01* BF_{10}=3.68	$t(26)$=0.97, p=0.34 BF_{10}=0.31

Note. Results of the independent two-sided t-tests for the mean of the predictors of main interest being equal to zero: expected value (EV) and expected risk (estimated when the trials' first number is presented), and reward prediction error (RPE) and surprise (estimated when the trial's reward or punishment are presented). These tests were run separately by regions of interest: left and right substantia nigra (SN), and left and right ventral tegmental area (VTA). Bayes factors (BF) higher than 1 provide evidence for an effect, while BF lower than 1 provide evidence for the absence of an effect.

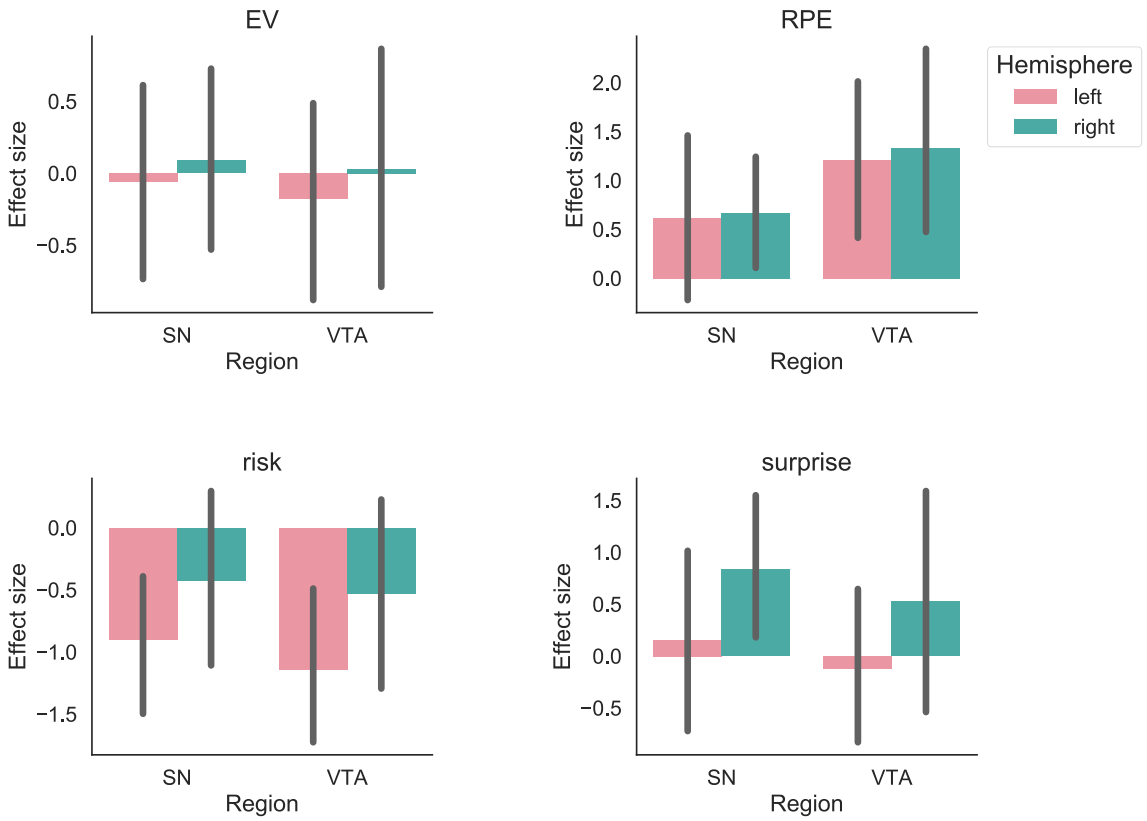


Figure 4.4: Average effect size across participants of the GLM on the time-series data extracted from the regions of interest (ROI): left and right substantia nigra (SN) and left and right ventral tegmental area (VTA). Different plots represent the predictors of main interest: expected value (EV) and expected risk (estimated when the trials' first number is presented), and reward prediction error (RPE) and surprise (estimated when the trial's reward or punishment are presented). Bars represent 95% confidence intervals.

and amygdala, and negative correlations with risk in anterior insula, dorsal striatum and posterior cingulate cortex. We found positive correlations with RPE in ventral striatum, precuneus, anterior insula and fusiform gyrus, and no negative correlations with RPE. Finally, we found positive correlations with surprise in the inferior frontal gyrus and superior temporal gyrus, and negative correlations with surprise in precuneus and posterior insula.

4.4 Discussion

Understanding the dopamine circuit is of great importance for both clinical and cognitive neuroscientists. First of all, the loss of dopaminergic neurons is associated with Parkinson's disease symptoms (Fearnley & Lees, 1991; Frank, 2006a) and dysregulations in the human dopamine circuit are known to play a role in drug addiction (Everitt & Robbins, 2005) and pathological gambling (Bergh, Eklund, Södersten, & Nordin, 1997). Moreover, the dopamine signal reflects different aspects of rewards: from the anticipation of risk to the mismatch between predictions and outcomes (Schultz, 2015). While dopamine neurons are situated mostly in the midbrain, they are part of a much greater and complex circuit, involving different cortical and subcortical areas (Watabe-Uchida et al., 2017; Haber & Knutson, 2010; Frank, 2006b). By transmitting information about changes in reward expectations and risk in the environment to areas important for action execution and learning, dopamine likely plays a crucial role in adaptive behavior, i.e., for survival in a dynamic environment, with limited resources and obstacles to avoid.

To date, most human studies have focused on the target areas (both cortical and subcortical) of the dopamine neurons because of methodological challenges. Importantly, human studies that investigated the activity of dopamine nuclei using fMRI provided incomplete and partially contradicting results. In this paper, we presented the results of a 7T fMRI study involving human participants performing a gambling task. To the best of our knowledge, this was the first study to investigate the functional role of both the VTA and the SN, using UHF-MRI to acquire high-quality, high-resolution functional and structural images. While previous studies in these areas focused on expected gains or losses and on the RPE signals, we extended the analysis to expected risk and to surprise. This was based on previous electrophysiological and fMRI studies that either found this signal in the VTA/SN or in their target areas (e.g., Fiorillo et al., 2003; Preuschoff et al., 2006; Hayden et al., 2011). While we found no evidence for a linear correlation between reward anticipation

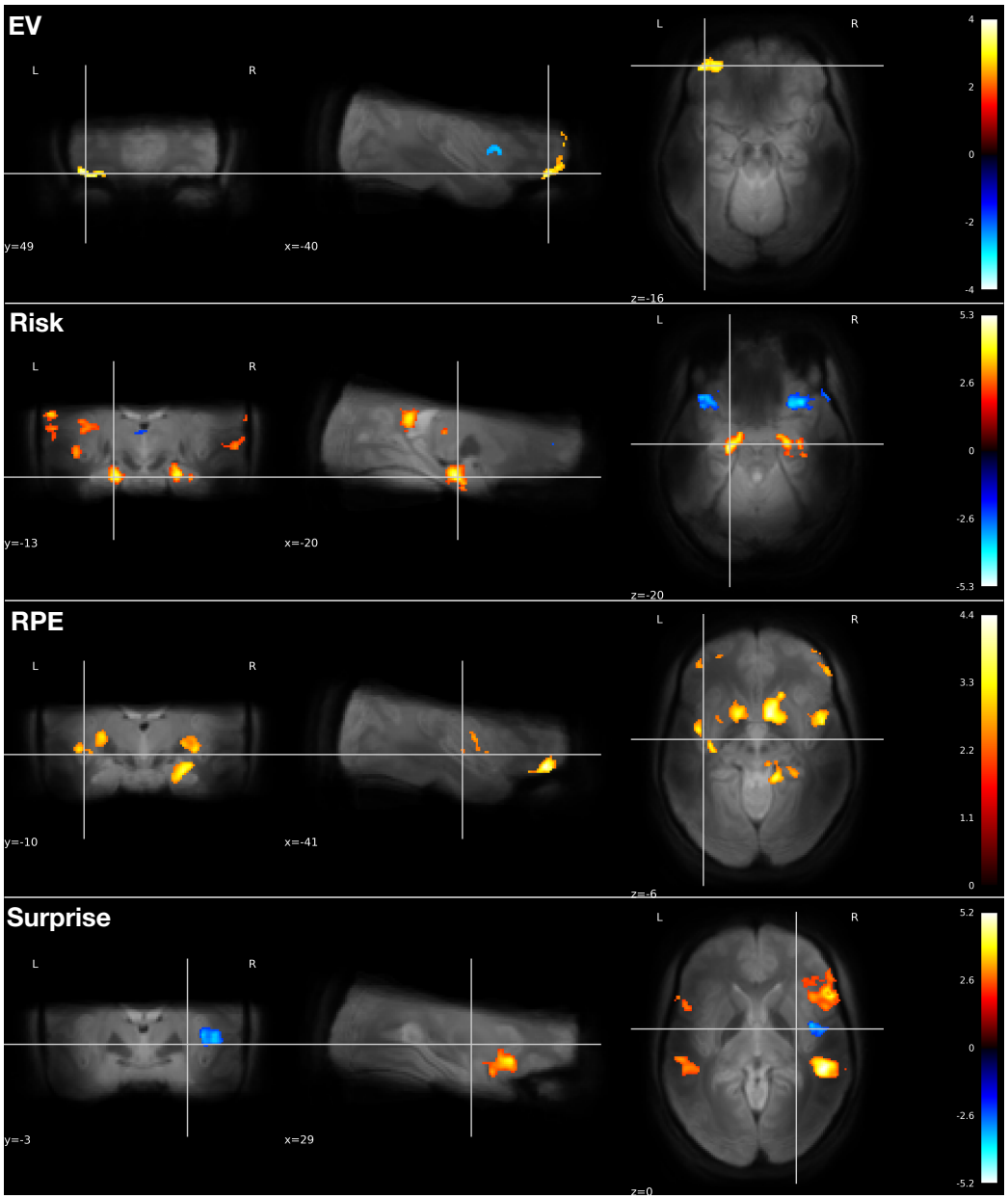


Figure 4.5: Results of the voxel-wise GLM after cluster correction, and overlapped onto the mean functional image across participants and volumes. Each row corresponds to the predictors of main interest: expected value (EV) and expected risk (estimated when the trials' first number is presented), and reward prediction error (RPE) and surprise (estimated when the trial's reward or punishment are presented).

Table 4.3: Results of the voxel-wise GLM after cluster-wise thresholding.

	Predictor	Cluster Index	Voxels	p	-log10(p)	Max	Max x (vox)	Max y (vox)	Max z (vox)
0	EV (positive)	2	939	<0.001	5.37	3.58	6.0	-54.0	18.0
1	EV (positive)	1	373	0.019	1.73	3.65	-42.0	49.5	-15.0
2	EV (negative)	4	1474	<0.001	8.15	-3.52	7.5	-28.5	0.0
3	EV (negative)	3	1082	<0.001	6.15	-3.96	15.0	-64.5	7.5
4	EV (negative)	2	565	0.001	3.09	-3.62	49.5	16.5	-1.5
5	EV (negative)	1	387	0.015	1.83	-3.38	-54.0	15.0	4.5
6	RPE (positive)	7	2409	<0.001	10.30	4.38	9.0	9.0	-4.5
7	RPE (positive)	6	1988	<0.001	8.73	3.89	-15.0	7.5	-9.0
8	RPE (positive)	5	1283	<0.001	5.90	3.98	25.5	-57.0	16.5
9	RPE (positive)	4	931	<0.001	4.29	4.17	43.5	9.0	-12.0
10	RPE (positive)	3	558	0.005	2.34	3.79	15.0	-39.0	-4.5
11	RPE (positive)	2	495	0.011	1.97	4.19	-40.5	46.5	-15.0
12	RPE (positive)	1	397	0.043	1.37	3.63	46.5	46.5	-13.5
13	risk (positive)	7	6993	<0.001	20.60	4.82	-52.5	-4.5	-9.0
14	risk (positive)	6	2859	<0.001	10.30	4.34	33.0	-40.5	7.5
15	risk (positive)	5	1122	<0.001	4.40	3.71	51.0	-33.0	19.5
16	risk (positive)	4	759	0.001	2.84	3.98	18.0	37.5	-6.0
17	risk (positive)	3	675	0.004	2.45	4.20	22.5	-13.5	-19.5
18	risk (positive)	2	522	0.02	1.69	3.59	69.0	1.5	13.5
19	risk (positive)	1	522	0.02	1.69	4.80	-22.5	-16.5	-19.5
20	risk (negative)	6	3931	<0.001	13.30	-5.30	33.0	22.5	-6.0
21	risk (negative)	5	1938	<0.001	7.22	-4.83	-25.5	19.5	-12.0
22	risk (negative)	4	807	0.001	3.06	-3.71	-36.0	52.5	3.0
23	risk (negative)	3	672	0.004	2.43	-3.82	-10.5	12.0	10.5
24	risk (negative)	2	570	0.012	1.94	-4.23	-3.0	-40.5	24.0
25	risk (negative)	1	451	0.048	1.32	-3.88	10.5	10.5	9.0
26	surprise (positive)	5	2639	<0.001	12.80	4.61	57.0	24.0	4.5
27	surprise (positive)	4	1723	<0.001	9.02	5.22	51.0	-34.5	0.0
28	surprise (positive)	3	1373	<0.001	7.22	3.83	-58.5	21.0	7.5
29	surprise (positive)	2	808	<0.001	4.44	3.99	-60.0	-52.5	9.0
30	surprise (positive)	1	760	<0.001	4.16	4.10	-54.0	-25.5	-9.0
31	surprise (negative)	2	572	0.001	3.00	-3.30	-10.5	-73.5	27.0
32	surprise (negative)	1	520	0.002	2.66	-3.77	40.5	-3.0	4.5

Note. Clusters that survive thresholding. We report the number of voxels, cluster probability, log probability, activation and MNI coordinate of the activation peak voxel in a cluster.

(involving both gains and losses) and VTA or SN activation, we did find evidence for a full RPE signal (positive and negative) in both regions, as well as for expected risk signal. Similarly to Matsumoto and Hikosaka (2009), who found a functional dissociation of VTA and SN, we also found a surprise signal in the SN but not in the VTA.

Both the SN and the VTA are relatively small brain areas (around 511 mm^3 and 138 mm^3, respectively, see Table 4.1), they are adjacent to each other as well as to other nuclei with related functions, such as the red nucleus and the subthalamic nucleus, and they are susceptible to other possible sources of noise, such as the physiological noise in the cerebrospinal fluid. The small dimension of the nuclei and their spatial contiguity increase the risk of confusing the signal from different regions (de Hollander et al., 2015). To be able to more reliably extract and separate the signals from the VTA and the SN, we therefore drew individual masks, based on 0.7 mm isotropic, multimodal, anatomical images that were acquired for each participant in a separate session. By restricting the analyses to the individual space, we also prevented disalignment issues that usually occur when transforming individual images to a group or standard space. To define the final masks, we adopted a rather conservative approach, by keeping the intersection of the masks drawn by two independent and trained raters. To illustrate the importance of these precautions, we compared our masks to previously proposed VTA and SN probabilistic masks in the standard space. In particular, we considered the SN subdivisions proposed by Zhang et al. (2017) and the VTA and the SN subdivisions proposed by Pauli et al. (2018). We found that, when transforming these masks to the individual space – as it is usually done during ROI signal extraction – the signal from the VTA and the SN is indeed partially confused. This can have serious impact on the interpretation of the results of an fMRI study. For instance, Zhang et al. (2017) reported a RPE signal in the medial part of the SN, which – according to our analyses and results – is the part that overlaps the most with the VTA, and a surprise signal in the lateral part of the SN. To be able to draw strong conclusions on the functional specificity of – in this case – SN subdivisions, it is preferable to have individually drawn masks.

Given previous findings (Fiorillo et al., 2003) and theoretical considerations (a reward-predicting cue can be seen as a RPE itself; see Hare, ODoherty, Camerer, Schultz, & Rangel, 2008), one might expect to find EV signals in the SN/VTA. However, as Berke (2018) recently discussed, EV signals are often found in dopamine target areas, such as the ventral striatum, but not in the dopamine cell firing itself (Eshel, Tian, Bukwich, & Uchida, 2016). Studies on the firing of dopaminergic nuclei are mostly done with animals that are

restricted from moving, therefore giving less emphasis on choices. In these tasks, animals are also extensively trained, to the point that conditioned stimuli – anticipating a reward or a punishment – are not surprising anymore. Therefore, the absence of an EV signal in both the SN and the VTA in our task is not too surprising: Since participants were explicitly instructed that the initial bet's outcome was random, there was also less focus on the action and more on the reward structure of the task. While we found positive correlation with EV in the orbital frontal cortex and negative correlation with EV in the anterior insula, in line with previous studies inspecting value signalling in the cortex (Schoenbaum, Takahashi, Liu, & McDannald, 2011; Bartra et al., 2013), the effect in ventral striatum did not survive the cluster thresholding.

The presence of a full RPE signal in both the VTA and the SN confirms previous results in animal studies (Schultz, 2015), although most of them are based on signal from the lateral part of the the VTA alone (Eshel et al., 2016). It also clarifies previous results on the VTA/SN signals in fMRI human studies (D'Ardenne et al., 2008; Pauli et al., 2015; Zhang et al., 2017). For instance, D'Ardenne et al. (2008) only found evidence for a positive – and not negative – RPE in VTA and not in SN. We also found an RPE signal in ventral striatum and anterior insula, confirming previous fMRI results that looked at dopamine target areas (Bartra et al., 2013).

Here, to the best of our knowledge, we showed for the first time the presence of a risk signal in both the VTA and the SN, in line with electrophysiological studies in non-human animals (Fiorillo et al., 2003). We also found risk signal in anterior cingulate cortex, amygdala and anterior insula, confirming previous fMRI studies linking these areas to the coding of risk (Preuschoff et al., 2006; J. W. Brown & Braver, 2018).

The presence of a surprise salience signal in the SN and not in the VTA is in line with results from the animal literature (Matsumoto & Hikosaka, 2009) and with the framework proposed by Bromberg-Martin et al. (2010). In this framework, there are two distinct functional groups of dopamine neurons, a motivational value group, that shows the standard RPE response, and a motivational salience group, that reflects how unexpected outcomes are – positive or negative alike. Cells of the first group are situated more in the dorsolateral part of the SNc, while cells of the second group are situated more in the ventromedial part of the SNc as well as in the VTA. While SNc cells project more to sensorimotor dorsolateral striatum, VTA cells project more to ventral striatum. Beyond our ROIs, we also found correlations between surprise and posterior (but not anterior) insula.

A limitation of our study is that we did not distinguish between the pars compacta and reticulata of the SN, while dopamine neurons are mainly situated in the pars compacta. However, these two parts are virtually indistinguishable based on MRI contrast alone (see Figure 4.2). Therefore, to avoid making an arbitrary decisions on where to set a border between the two, we considered the SN as one structure. By combining different methodologies (i.e., diffusion MRI) further studies might shed further light on SN functional subdivisions.

Since the BOLD response measured in fMRI is an indirect measure of neuronal activity and is mainly thought to measure signals input and local processing of neurons rather than their output (Logothetis & Wandell, 2004), it is important to integrate results from different methodologies and species in order to understand the complexity of the dopaminergic circuit as a whole.

In sum, in this study we used novel methodologies to investigate how the brain processes gains and losses and updates expectations based on experience. We were able to show for the first time a risk signal in the dopamine nuclei with human participants, and provided evidence for a full RPE signal in the presence of both gains and losses, thus clarifying previous results of human fMRI studies. This study opens the way to a better understanding of the dopamine circuit in the human brain, especially regarding the functional specificity of the SN and the VTA (or of their subregions) in reward-based decision making and adaptive behavior. Future studies might extend these methodologies to different paradigms in which participants need to more actively interact with the environment, as in trial-and-error learning (Sutton & Barto, 1998).

References

Andersson, J. L. R., Skare, S., & Ashburner, J. (2003). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *NeuroImage*, *20*(2), 870–888. doi: 10.1016/S1053-8119(03)00336-7

Arias-Carrión, O., Stamelou, M., Murillo-Rodríguez, E., Menéndez-Gonzáles, M., & Pöppel, E. (2010). Dopaminergic reward system: A short integrative review. *International Archives of Medicine*, *3*(24), 1–6. doi: 10.1186/1755-7682-3-24

Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, *12*(1), 26–41. doi: 10.1016/j.media.2007.06.004

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate-based meta-analysis of bold fmri experiments examining neural correlates of subjective value. *NeuroImage*, *76*, 412–427. doi: 10.1016/j.neuroimage.2013.02.063

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, *50*(1-3), 7–15. doi: 10.1016/0010-0277(94)90018-3

Behzadi, Y., Restom, K., Liau, J., & T.Liu, T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, *37*(1), 90–101. doi: 10.1016/j.neuroimage.2007.04.042

Bergh, C., Eklund, T., Södersten, P., & Nordin, C. (1997). Altered dopamine function in pathological gambling. *Psychological Medicine*, *27*(2), 473–475. doi: 10.1017/S0033291796003789

Berke, J. D. (2018). What does dopamine mean? *Nature Neuroscience*, *21*, 787–793. doi: 10.1038/s41593-018-0152-y

Boehm, U., Annis, J., Frank, M. J., Hawkins, G. E., Heathcote, A., Kellen, D., ... Wagenmakers, E.-J. (2018). Estimating across-trial variability parameters of the diffusion decision model: Expert advice and recommendations. *Journal of Mathematical Psychology*, *28*, 46–75. doi: 10.1016/j.jmp.2018.09.004

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700–765. doi: 10.1037/0033-295X.113.4.700

Bogacz, R., Wagenmakers, E.-J., Forstmann, B. U., & Nieuwenhuis, S. (2010). The neural basis of the speedaccuracy tradeoff. *Trends in Neuroscience*, *33*(1), 10–16. doi: 10.1016/j.tins.2009.09.002

Boureau, Y.-L., & Dayan, P. (2011). Opponency revisited: Competition and cooperation between dopamine and serotonin. *Neuropsychopharmacology*, *36*(1), 74. doi: 10.1038/npp.2010

Bridle, J. S. (1990). Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimates of parameters. In D. S. Touretzky (Ed.), *Advances in neural information processing systems: Proceedings of the 1989 conference* (pp. 211–217). Morgan Kaufman, San Mateo, CA.

Bromberg-Martin, E. S., Matsumoto, M., & Hikosaka, O. (2010). Dopamine in motivational control: Rewarding, aversive, and alerting. *Neuron*, *68*(5), 815–834. doi: 10.1016/j.neuron.2010.11.022

Brown, J. W., & Braver, T. (2018). Risk prediction and aversion by anterior cingulate cortex. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(4), 266–277. doi: 10.3758/CABN.7.4.266

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*, 153–178.

Bürkner, P.-C. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, *80*(1), 1–28. doi: 10.18637/jss.v080.i01

Busemeyer, J. R., Stout, J. C., & Finn, P. R. (2003). Using computational models to help explain decision making: Processes of substance abusers. In D. Barch (Ed.), *Cognitive and affective neuroscience of psychopathology.* New York: Oxford University Press.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological review*, *100*(3), 432. doi: 10.1037/0033-295X.100.3.432

Busemeyer, J. R., Wang, Z., Townsend, J. T., & Eidels, A. (2015). *The Oxford handbook of computational and mathematical psychology.* New York, NY: Oxford University Press. doi: 10.1111/bjop.12201

Caan, M., Bazin, P.-L., Fracasso, A., Marques, J., Dumoulin, S., & van der Zwaag, W. (2018). *MP2RAGEME: T_1, T_2^* and QSM mapping in one sequence at 7 Tesla.* (Poster presented at the Joint Annual Meeting ISMRM-ESMRMB, Paris, France)

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. doi: 10.18637/jss.v076.i01

Cavanagh, J. F., Wiecki, T. V., Cohen, M. X., Figueroa, C. M., Samanta, J., Sherman, S. J., & Frank, M. J. (2011). Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nature neuroscience*, *14*(11), 1462. doi: 10.1038/nn.2925

Cavanagh, J. F., Wiecki, T. V., Kochar, A., & Frank, M. J. (2014). Eye tracking and pupillometry are indicators of dissociable latent decision processes. *Journal of Experimental Psychology: General*, *143*(4), 1476–1488. doi: 10.1037/a0035813

Christakou, A., Gershman, S. J., Niv, Y., Simmons, A., Brammer, M., & Rubia, K. (2013). Neural and psychological maturation of decision-making in adolescence and young adulthood. *Journal of cognitive neuroscience*, *25*(11), 1807–1823. doi: 10.1162/jocn_a_00447

Clithero, J. A. (2018). Improving out-of-sample predictions using response times and a model of the decision process. *Journal of Economic Behavior & Organization*, *148*, 344–375. doi: 10.1016/j.jebo.2018.02.007

Clithero, J. A., & Rangel, A. (2014). Informatic parcellation of the network involved in the computation of subjective value. *Social Cognitive and Affective Neuroscience*, *9*(9), 1289–1302. doi: 10.1093/scan/nst106

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, *29*(3), 162–173. doi: 10.1006/cbmr.1996.0014

D'Ardenne, K., McClure, S. M., Nystrom, L. E., & Cohen, J. D. (2008). BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. *Science*, *319*(5867), 1264–1267. doi: 10.1126/science.1150605

Dayan, P., & Abbott, L. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems.* Cambridge, MA: MIT press.

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(4), 429–453. doi: 10.3758/CABN.8.4.42

de Hollander, G., Keuken, M. C., & Forstmann, B. U. (2015). The subcortical cocktail problem; mixed signals from the subthalamic nucleus and substantia nigra. *PloS One*, *10*(2), 1–18. doi: 10.1371/journal.pone.0120572

de Hollander, G., Keuken, M. C., van der Zwaag, W., Forstmann, B. U., & Trampel, R. (2017). Comparing functional MRI protocols for small, iron-rich basal ganglia nuclei such as the subthalamic nucleus at 7T and 3T. *Human Brain Mapping*, *38*(6), 3226–3248. doi: 10.1002/hbm.23586

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*(3), 297–302. doi: 10.2307/1932409

Diederen, K. M., & Schultz, W. (2015). Scaling prediction errors to reward variability benefits error-driven learning in humans. *Journal of Neurophysiology*, *114*(3), 1628–1640.

Donner, T. H., Siegel, M., Fries, P., & Engel, A. K. (2009). Buildup of choice-predictive activity in human motor cortex during perceptual decision making. *Current Biology*, *19*, 1581–1585. doi: 10.1016/j.cub.2009.07.066

Dutilh, G., & Rieskamp, J. (2016). Comparing perceptual and preferential decision making. *Psychonomic Bulletin & Review*, 723–737. doi: 10.3758/s13423-015-0941-1

Dutilh, G., van Ravenzwaaij, D., Nieuwenhuis, S., van der Maas, H. L., Forstmann, B. U., & Wagenmakers, E.-J. (2012). How to measure post-error slowing: A confound and a simple solution. *Journal of Mathematical Psychology*(56), 208–2016. doi: 10.1016/j.jmp.2012.04.001

Eapen, M., Zald, D. H., Gatenby, J. C., Ding, Z., & Gore, J. C. (2011). Using high-resolution MR imaging at 7T to evaluate the anatomy of the midbrain dopaminergic system. *American Journal of Neuroradiology*, *32*(4), 688–694. doi: 10.3174/ajnr.A2355

Erev, I. (1998). Signal detection by human observers: A cutoff reinforcement learning model of categorization decisions under uncertainty. *Psychological Review*, *105*(2), 280–298. doi: 10.1037/0033-295X.105.2.280

Eshel, N., Tian, J., Bukwich, M., & Uchida, N. (2016). Dopamine neurons share common response function for reward prediction error. *Nature Neuroscience*, *19*, 479–486. doi: 10.3174/ajnr.A2355

Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., & Gorgolewski, K. J. (2017). MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PloS One*, *12*(9), 1–21. doi: 10.1371/journal.pone.0184661

Esteban, O., Markiewicz, C., Blair, R. W., Moodie, C., Isik, A. I., Aliaga, A. E., ... Gorgolewski, K. J. (2018). *FMRIPrep: a robust preprocessing pipeline for functional MRI.* doi: 10.1101/306951

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review*, *57*(2), 94–107. doi: 10.1037/h0058559

Everitt, B. J., & Robbins, T. W. (2005). Neural systems of reinforcement for drug addiction: From actions to habits to compulsion. *Nature Neuroscience*, *8*(11), 1481–1489. doi: 10.1038/nn1579

Fearnley, J. M., & Lees, A. J. (1991). Ageing and Parkinson's disease: Substantia nigra regional selectivity. *Brain*, *114*(5), 2283–2301. doi: 10.1093/brain/114.5.2283

Fiorillo, C. D., Newsome, W. T., & Schultz, W. (2008). The temporal precision of reward prediction in dopamine neurons. *Nature neuroscience*, *11*(8), 966. doi: 10.1038/nn.2159

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, *299*(5614), 1898–1902. doi: 10.1126/science.1077349

Fonov, V. S., Evans, A. C., McKinstry, R. C., Almli, C. R., & Collins, D. L. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, *4*(Supplement 1), 39–41. doi: 10.1016/S1053-8119(09)70884-5

Forstmann, B. U., Anwander, A., Schäfer, A., Neumann, J., Brown, S. D., Wagenmakers, E.-J., ... Turner, R. (2010). Cortico-striatal connections predict control over speed and accuracy in perceptual decision making. *PNAS*, *107*(36), 15916–15920. doi: 10.1073/pnas.1004932107

Forstmann, B. U., Dutilh, G., Brown, S. D., Neumann, J., von Cramon, D. Y., Ridderinkhof, K. R., & Wagenmakers, E.-J. (2008). Striatum and pre-sma facilitate decision-making under time pressure. *PNAS*, *105*(45), 17538–17542. doi: 10.1073/pnas.0805903105

Forstmann, B. U., Tittgemeyer, M., Wagenmakers, E.-J., Derrfuss, J., Imperati, D., & Brown, S. D. (2011). The speed-accuracy tradeoff in the elderly brain: A structural model-based approach. *The Journal of Neuroscience*, *31*(47), 17242–17249. doi: 10.1523/jneurosci.0309-11.2011

Forstmann, B. U., & Wagenmakers, E.-J. (2015). *An introduction to model-based cognitive neuroscience*. Springer. doi: 10.1007/978-1-4939-2236-9

Fouragnan, E., Retzler, C., & Philiastides, M. G. (2018). Separate neural representations of prediction error valence and surprise: Evidence from an fMRI meta-analysis. *Human Brain Mapping*, *39*, 2887–2906. doi: 10.1002/hbm.24047

Frank, M. J. (2006a). Dynamic dopamine modulation in the basal ganglia: A neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. *Journal of Cognitive Neuroscience*, *17*(1), 17–52. doi: 10.1162/0898929052880093

Frank, M. J. (2006b). Hold your horses: a dynamic computational role for the subthalamic nucleus in decision making. *Neural Networks*, *19*(8), 1120–1136.

Frank, M. J., Gagne, C., Nyhus, E., Masters, S., Wiecki, T. V., & Badre, D. (2015). fMRI and EEG predictors of dynamic decision parameters during human reinforcement learning. *The Journal of Neuroscience*, *35*(2), 485–494. doi: 10.1523/JNEUROSCI.2036-14.2015

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences*, *104*(41), 16311–16316. doi: 10.1073/pnas.0706111104

Frank, M. J., Samanta, J., Moustafa, A. A., & Sherman, S. J. (2007). Hold your horses: Impulsivity, deep brain stimulation, and medication in parkinsonism. *Science*, *318*, 1309–1312. doi: 10.1126/science.1146157

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By carrot or by stick: Cognitive reinforcement learning in parkinsonism. *Science*, *306*(5703), 1940–1943. doi: 10.1126/science.1102941

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis.* (3rd ed.). Chapman & Hall/CRC.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*(4), 733–807.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. doi: 10.1214/ss/1177011136

Gershman, S. J. (2015). Do learning rates adapt to the distribution of rewards? *Psychonomic Bulletin & Review*, *22*, 1320–1327. doi: 10.3758/s13423-014-0790-3

Glimcher, P. W., & Fehr, E. (2014). *Neuroeconomics. Decision making and the brain.* (2nd

ed.). Academic Press. doi: 10.1016/C2011-0-05512-6

Gluth, S., & Rieskamp, J. (2017). Variability in behavior that cognitive models do not explain can be linked to neuroimaging data. *Journal of Mathematical Psychology*, *76*, 104–116. doi: 10.1016/j.jmp.2016.04.012

Gluth, S., Rieskamp, J., & Büchel, C. (2012). Deciding when to decide: Time-variant sequential sampling models explain the emergence of value-based decisions in the human brain. *Journal of Neuroscience*, *32*(31), 10686–10698. doi: 10.1523/JNEUROSCI.0727-12.2012

Gluth, S., Rieskamp, J., & Büchel, C. (2013). Classic EEG motor potentials track the emergence of value-based decisions. *Neuroimage*, *79*(1), 394–403. doi: 10.1016/j.neuroimage.2013.05.005

Gold, J. I., & Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences*, *5*(1), 10–16. doi: 10.1016/S1364-6613(00)01567-9

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*, 535–574. doi: 10.1146/annurev.neuro.29.051605.113038

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in Python. *Frontiers in Neuroinformatics*, *5*(13), 1–15. doi: 10.3389/fninf.2011.00013

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* Oxford, England: John Wiley.

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, *48*(1), 63–72. doi: 10.1016/j.neuroimage.2009.06.060

Gu, R., Feng, X., Broster, L. S., Yuan, L., Xu, P., & Luo, Y.-j. (2017). Valence and magnitude ambiguity in feedback processing. *Brain and behavior*, *7*(5), e00672. doi: 10.1002/brb3.672

Haber, S. N., & Knutson, B. (2010). The reward circuit: Linking primate anatomy and human imaging. *Neuropsychopharmacology*, *35*(1), 4–26. doi: 10.1038/npp.2009.129

Hanks, T. D., Ditterich, J., & Shadlen, M. N. (2006). Microstimulation of macaque area LIP affects decision-making in a motion discrimination task. *Nature Neuroscience*, *9*(5), 682–9. doi: 10.1038/nn1683

Hanks, T. D., & Summerfield, C. (2017). Perceptual decision making in rodents, monkeys, and humans. *Neuron*, *93*(1), 15–31. doi: 10.1016/j.neuron.2016.12.003

Hare, T. A., ODoherty, J., Camerer, C. F., Schultz, W., & Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *The Journal of Neuroscience*, *28*(22), 5623–5630. doi: 10.1523/JNEUROSCI.1309-08.2008

Hayden, B. Y., Heilbronner, S. R., Pearson, J. M., & Platt, M. L. (2011). Surprise signals in anterior cingulate cortex: Neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. *Journal of Neuroscience*, *31*(11), 4178–4187. doi: 10.1523/JNEUROSCI.4652-10.2011

Heitz, R. P. (2008). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *9*(150), 467–479. doi: 10.1038/nrn2374

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*(4), 679–709. doi: 10.1037/0033-295X.109.4.679

Huk, A., & Meister, M. L. R. (2012). Neural correlates and neural computations in posterior parietal cortex during perceptual decision-making. *Frontiers in Integrative Neuroscience*, *6*(86), 1–13. doi: 10.3389/fnint.2012.00086

Hunt, L. T., Kolling, N., Soltani, A., Woolrich, M. W., Rushworth, M. F., & Behrens, T. E. (2012). Mechanisms underlying cortical activity during value-guided choice. *Nature Neuroscience*, *15*(3), 470–S3. doi: 10.1038/nn.3017

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 computational biology*, *7*(4), e1002028. doi: 10.1371/journal.pcbi.1002028

Jarosz, A. F., & Wiley, J. (2014). What are the odds? a practical guide to computing and reporting bayes factors. *Journal of Problem Solving*, *7*(1), 2–9. doi: 10.7771/1932-6246.1167

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, *17*(2), 825–841. doi: 10.1006/nimg.2002.1132

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–292. doi: 10.2307/1914185

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Katz, L. N., Yates, J. L., Pillow, J. W., & Huk, A. C. (2016). Dissociated functional significance of decision-related activity in the primate dorsal stream. *Nature*, *535*, 285–288. doi: 10.1038/nature18617

Kelly, S. P., & O'Connell, R. G. (2013). Internal and external influences on the rate of sensory evidence accumulation in the human brain. *Journal of Neuroscience*, *33*(50), 19434–19441. doi: 10.1523/JNEUROSCI.3355-13.2013

Keuken, M. C., Bazin, P.-L., Crown, L., Hootsmans, J., Laufer, A., Mller-Axt, C., ... Forstmann, B. U. (2014). Quantifying inter-individual anatomical variability in the subcortex using 7 T structural MRI. *NeuroImage*, *94*(1), 40–46. doi: 10.1016/j .neuroimage.2014.03.032

Keuken, M. C., & Forstmann, B. U. (2015). A probabilistic atlas of the basal ganglia using 7 t mri. *Data in brief*, *4*, 577–582. doi: 10.1016/j.dib.2015.07.028

Kocher, M. G., & Sutter, M. (2006). Time is money – time pressure, incentives, and the quality of decision-making. *Journal of Economic Behavior & Organization*, *61*(3), 375–392.

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Natire Neuroscience*, *13*, 1292–1298. doi: 10.1038/nn.2635

Krajbich, I., Lu, D., Camerer, C., & Rangel, A. (2012). The attentional drift-diffusion model extends to simple purchasing decisions. *Frontiers in Psychology*, *3*, 1–18. doi: 10.3389/fpsyg.2012.00193

LaBerge, D. A. (1994). Quantitative models of attention and response processes in shape identification tasks. *Journal of Mathematical Psychology*, *38*, 198-243. doi: 10.1006/ jmps.1994.1015

Langkammer, C., Schweser, F., Krebs, N., Deistung, A., Goessler, W., Scheurer, E., ... Reichenbach, J. R. (2012). Quantitative susceptibility mapping (QSM) as a means to measure brain iron? A post mortem validation study. *NeuroImage*, *62*(3), 1593–1599. doi: 10.1016/j.neuroimage.2012.05.049

Lebreton, M., Bacily, K., Palminteri, S., & Engelmann, J. B. (2018). Contextual influence on confidence judgments in human reinforcement learning. *bioRxiv*. doi: 10.1101/339382

Lebreton, M., Jorge, S., Michel, V., Thirion, B., & Pessiglione, M. (2009). An automatic val-

uation system in the human brain: Evidence from functional neuroimaging. *Neuron*, *64*(3), 431–439. doi: 10.1016/j.neuron.2009.09.040

Lebreton, M., Langdon, S., Slieker, M. J., Nooitgedacht, J. S., Goudriaan, A. E., Denys, D., ... Luigjes, J. (2018). Two sides of the same coin: Monetary incentives concurrently improve and bias confidence judgments. *Science Advances*, *4*(5), eaaq0668. doi: 10.1126/sciadv.aaq0668

Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., & Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, *1*(67), 1–10. doi: 10.1038/s41562-017-0067

Leite, F., & Ratcliff, R. (2011). What cognitive processes drive response biases? a diffusion model analysis. *Journal of Mathematical Psychology*, *36*(7), 651–687.

Lewandowsky, S., & Simon, F. (2010). *Computational modeling in cognition: Principles and practice.* Sage Publications.

Logothetis, N. K., & Wandell, B. A. (2004). Interpreting the BOLD signal. *Annual Review of Physiology*, *66*, 735–69. doi: 10.1146/annurev.physiol.66.082602.092845

Luce, R. D. (1959). *Individual choice behavior.* New York: John Wiley and Sons.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization.* Oxford University Press.

Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature neuroscience*, *14*(2), 154. doi: 10.1038/nn.2723

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, *7*(1), 77–91. doi: 10.1111/j.1540-6261.1952.tb01525.x

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* New York: Henry Holt and Co., Inc.

Matsumoto, M., & Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature*, *459*(7248), 837–841. doi: 10.1038/nature08028

Metere, R., Kober, T., Möller, H. E., & Schäfer, A. (2017). Simultaneous quantitative MRI mapping of T_1, T_2^* and magnetic susceptibility with multi-echo MP2RAGE. *PloS One*, *12*(1), 1–28. doi: 10.1371/journal.pone.0169265

Milosavljevic, M., Malmaud, J., Huth, A., Koch, C., & Rangel, A. (2010). The drift diffusion model can account for the accuracy and reaction time of value-based choices

under high and low time pressure. *Judgment and Decision Making*, *5*(6), 437–449. doi: 10.2139/ssrn.1901533

Morey, R. D., Rouder, J. N., & Jamil, T. (2015). Bayesfactor: Computation of bayes factors for common designs [Computer software manual]. (R package version 0.9)

Mulder, M. J., VanMaanen, L., & Forstmann, B. U. (2014). Perceptual decision neuroscience: A model-based review. *Neuroscience*, *277*, 872–884. doi: 10.1016/j.neuroscience.2014.07.031

Mulder, M. J., Wagenmakers, E. J., Ratcliff, R., Boekel, W., & Forstmann, B. U. (2012). Bias in the brain: a diffusion model analysis of prior probability and potential payoff. *Journal of Neuroscience*, *32*(7), 2335–2343. doi: 10.1523/JNEUROSCI.4156-11.2012

Navarro, D. J., & G.Fuss, I. (2009). Fast and accurate calculations for first-passage times in wiener diffusion models. *Journal of Mathematical Psychology*, *53*(4), 222–230. doi: 10.1016/j.jmp.2009.02.003

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*, 139–154. doi: 10.2307/1914185

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. *Journal of Neuroscience*, *32*(2), 551–562. doi: 10.1523/JNEUROSCI.5498-10.2012

Niv, Y., & Schoenbaum, G. (2008). Dialogues on prediction errors. *Trends in cognitive sciences*, *12*(7), 265–272. doi: 10.1016/j.tics.2008.03.006

O'Connell, R. G., Dockree, P. M., & Kelly, S. P. (2012). A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nature Neuroscience*, *15*, 1729–1735. doi: 10.1038/nn.3248

O'Doherty, J. P. (2004). Reward representations and reward-related learning in the human brain: Insights from neuroimaging. *Current Opinion in Neurobiology*, *14*(6), 769–776. doi: 10.1016/j.conb.2004.10.016

O'Doherty, J. P., & Bossaerts, P. (2008). Toward a mechanistic understanding of human decision making: Contributions of functional neuroimaging. *Current Directions in Psychological Science*, *17*(2), 119–123. doi: 10.1111/j.1467-8721.2008.00560.x

O'Doherty, J. P., Hampton, A., & Kim, H. (2007). Model-based fmri and its application to reward learning and decision making. *Annals of the New York Academy of Sciences*, *1104*, 35–53. doi: 10.1196/annals.1390.022

Oud, B., Krajbich, I., Miller, K., Cheong, J. H., Botvinick, M., & Fehr, E. (2016). Irrational

time allocation in decision-making. *Proceedings of the Royal Society B*, *283*(1822), 1–8. doi: 10.1098/rspb.2015.1439

Palminteri, S., Khamassi, M., Joffily, M., & Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning. *Nature Communications*, *6*(8096), 1–14. doi: 10.1038/ncomms9096

Palminteri, S., Kilford, E. J., Coricelli, G., & Blakemore, S.-J. (2016). The computational development of reinforcement learning during adolescence. *PLOS Computational Biology*, *12*(e1004953). doi: 10.1371/journal.pcbi.1004953

Palminteri, S., Lefebvre, G., Kilford, E. J., & Blakemore, S.-J. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLOS Computational Biology*, *13*(e1005684). doi: 10.1371/journal.pcbi.1005684

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in Cognitive Science*, *21*(6), 25–433. doi: 10.1016/j.tics.2017.03.011

Pauli, W. M., Larsen, T., Collette, S., Tyszka, J. M., Seymour, B., & O'Doherty, J. P. (2015). Distinct contributions of ventromedial and dorsolateral subregions of the human substantia nigra to appetitive and aversive learning. *The Journal of Neuroscience*, *13*(42), 14220–14233. doi: 10.1523/JNEUROSCI.2277-15.2015

Pauli, W. M., Nili, A. N., & Tyszka, J. M. (2018). A high-resolution probabilistic in vivo atlas of human subcortical brain nuclei. *Scientific Data*, *5*(180063), 1–13. doi: 10.1038/sdata.2018.63

Pavlov, I. P. (1927). *Conditional reflexes: An investigation of the physiological activity of the cerebral cortex.* Oxford University Press.

Pearce, J. M., & Hall, G. (1980). A model for pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological review*, *87*(6), 532.

Pedersen, M. L., Frank, M. J., & Biele, G. (2017). The drift diffusion model as the choice rule in reinforcement learning. *Psychonomic Bulletin & Review*, *24*(4), 1234-1251. doi: 10.3758/s13423-016-1199-y

Peirce, J. W. (2007). Psychopypsychophysics software in python. *Journal of Neuroscience Methods*, *162*(1-2), 8–13. doi: 10.1016/j.jneumeth.2006.11.017

Pirrone, A., Azab, H., Hayden, B. Y., Stafford, T., & Marshall, J. A. R. (2017). Evidence for the speedvalue trade-off: Human and monkey decision making is magnitude sensitive.

Decision. doi: 10.1037/dec0000075

Polania, R., Krajbich, I., Grueschow, M., & Ruff, C. C. (2014). Neural oscillations and synchronization differentially support evidence accumulation in perceptual and value-based decision making. *Neuron*, *82*, 709–720. doi: 10.1016/j.neuron.2014.03.014

Polania, R., Moisa, M., Opitz, A., Grueschow, M., & Ruff, C. C. (2015). The precision of value-based choices depends causally on fronto-parietal phase coupling. *Nature communications*, *6*, 8090. doi: 10.1038/ncomms9090(2015)

Power, J. D., Mitra, A., Laumann, T. O., Z.Snyder, A., L.Schlaggar, B., & E.Petersen, S. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage*, *84*(1), 320–341. doi: 10.1016/j.neuroimage.2013.08.048

Preuschoff, K., Bossaerts, P., & Quartz, S. R. (2006). Neural differentiation of expected reward and risk in human subcortical structures. *Neuron*, *51*, 381–390. doi: 10.1016/j.neuron.2006.06.024

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. doi: 10.1037/0033-295X.85.2.59

Ratcliff, R. (1985). Theoretical interpretations of the speed and accuracy of positive and negative responses. *Psychological Review*, *92*(2), 212–225. doi: 10.1037/0033-295X.92.2.212

Ratcliff, R., & Frank, M. J. (2012). Reinforcement-based decision making in corticostriatal circuits: mutual constraints by neurocomputational and diffusion models. *Neural computation*, *24*(5), 1186–1229. doi: 10.1162/NECO_a_00270

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356. doi: 10.1111/1467-9280.00067

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, *111*(2), 333–367. doi: 10.1037/0033-295X.111.2.333

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *61*(6), 260–281. doi: 10.1016/j.tics.2016.01.007

Ratcliff, R., Thapar, A., & Mckoon, G. (2003). A diffusion model analysis of the effects of aging on brightness discrimination. *Perception & Psychophysics*, *65*(4), 523–535. doi: /10.3758/BF03194580

Ratcliff, R., Voskuilen, C., & Teodorescu, A. (2018). Modeling 2-alternative forced-choice

tasks: Accounting for both magnitude and difference effects. *Cognitive psychology*, *103*, 1–22. doi: 10.1016/j.cogpsych.2018.02.002

Rescorla, R., & Wagner, A. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. Black & W. Prokasy (Eds.), *Classical conditioning ii: Current research and theory* (pp. 64–99). Appleton-Century-Crofts.

Rieskamp, J., & Otto, P. E. (2006). Ssl: a theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*(2), 207–236. doi: 10.1037/0096-3445.135.2.207

Salvador, A., Worbe, Y., Delorme, C., Coricelli, G., Gaillard, R., Robbins, T. W., ... Palminteri, S. (2017). Specific effect of a dopamine partial agonist on counterfactual learning: evidence from Gilles de la Tourette syndrome. *Scientific reports*, *7*(1), 6292. doi: 10.1038/s41598-017-06547-8

Schoenbaum, G., Takahashi, Y., Liu, T.-L., & McDannald, M. A. (2011). Does the orbitofrontal cortex signal value? *Annals of the New York Academy of Sciences*, *1239*, 87–99. doi: 10.1111/j.1749-6632.2011.06210.x

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of neurophysiology*, *80*(1), 1–27. doi: 10.1152/jn.1998.80.1.1

Schultz, W. (2010). Dopamine signals for reward value and risk: basic and recent data. *Behavioral and brain functions*, *6*(1), 24. doi: 10.1186/1744-9081-6-24

Schultz, W. (2015). Neuronal reward and decision signals: From theories to data. *Physiological Reviews*, *95*, 853–951. doi: 10.1152/physrev.00023.2014

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599. doi: 10.1126/science.275.5306.1593

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In *Proceedings of the 9th Python in science conference* (pp. 57–61).

Shenhav, A., Straccia, M. A., Cohen, J. D., & Botvinick, M. M. (2014). Anterior cingulate engagement in a foraging context reflects choice difficulty, not foraging value. *Nature neuroscience*, *17*, 1249–1254. doi: 10.1038/nn.3771

Singer, T., Critchley, H. D., & Preuschoff, K. (2009). A common role of insula in feelings, empathy and uncertainty. *Trends in Cognitive Sciences*, *13*(8), 334–340. doi: 10.1016/j.tics.2009.05.001

Singmann, H., Klauer, K. C., & Kellen, D. (2014). Intuitive logic revisited: New data

and a bayesian mixed model meta-analysis. *PLOS one*, *9*(4), e94223. doi: 10.1371/journal.pone.0094223

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis.* Oxford, England: Appleton-Century.

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, *27*(3), 161–168. doi: 10.1016/j.tins.2004.01.006

Spektor, M. S., Gluth, S., Fontanesi, L., & Rieskamp, J. (in press). How similarity between choice options affects decisions from experience: The accentuation of differences model. *Psychological Review*. doi: 10.1037/rev0000122

Spektor, M. S., & Kellen, D. (2018). The relative merit of empirical priors in non-identifiable and sloppy models: Applications to models of learning and decision-making. *Psychonomic Bulletin & Review*. doi: 10.3758/s13423-018-1446-5

Steinberg, E. E., Keiflin, R., Boivin, J. R., Witten, I. B., Deisseroth, K., & Janak, P. H. (2013). A causal link between prediction errors, dopamine neurons and learning. *Nature neuroscience*, *16*(7), 966. doi: 10.1038/nn.3413

Steingroever, H., Wetzels, R., & Wagenmakers, E.-J. (2013). A comparison of reinforcement learning models for the iowa gambling task using parameter space partitioning. *The Journal of Problem Solving*, *5*(2), 2.

Steingroever, H., Wetzels, R., & Wagenmakers, E.-J. (2014). Absolute performance of reinforcement-learning models for the iowa gambling task. *Decision*, *1*(3), 161.

Summerfield, C., & Tsetsos, K. (2012). Building bridges between perceptual and economic decision-making: Neural and computational mechanisms. *Frontiers in Neuroscience*, *6*(70), 1–20. doi: 10.3389/fnins.2012.00070

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction.* Cambridge, MA: MIT press.

Teodorescu, A. R., Moran, R., & Usher, M. (2015). Absolutely relative or relatively absolute: Violations of value invariance in human decision making. *Psychonomic Bulletin & Review*, *23*(1), 22–38. doi: 10.3758/s13423-015-0858-8

Thorndike, E. L. (1911). *Animal intelligence: Experimental studies.* New York: The Macmillan Company.

Turner, B. M., Schley, D. R., Muller, C., & Tsetsos, K. (2018). Competing theories of multialternative, multiattribute preferential choice. *Psychological Review*, *125*(3), 329–362. doi: 10.1037/rev0000089

Turner, B. M., van Maanen, L., & Forstmann, B. U. (2015). Informing cognitive abstractions through neuroimaging: the neural drift diffusion model. *Psychological Review*, *122*(2), 312–336. doi: 10.1037/a0038894

Turner, B. M., Wang, T., & C.Merkle, E. (2017). Factor analysis linking functions for simultaneously modeling neural and behavioral data. *NeuroImage*, *153*, 28–48. doi: 10.1016/j.neuroimage.2017.03.044

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: improved N3 bias correction. *IEEE Transactions on Medical Imaging*, *29*(6), 1310–1320. doi: 10.1109/TMI.2010.2046908

Usher, M., & McClelland, J. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550–592. doi: 10.1037/0033-295X.108.3.550

van der Zwaag, W., Schäfer, A., Marques, J. P., Turner, R., & Trampel, R. (2015). Recent applications of UHF-MRI in the study of human brain function and structure: A review. *NMR in Biomedicine*, *29*(9), 1274–1288. doi: 10.1002/nbm.3275

van Maanen, L., Brown, S. D., Eichele, T., Wagenmakers, E.-J., Ho, T., Serences, J., & Forstmann, B. U. (2011). Neural correlates of trial-to-trial fluctuations in response caution. *Journal of Neuroscience*, *31*(48), 17488–17495. doi: 10.1523/JNEUROSCI.2924-11.2011

van Maanen, L., Fontanesi, L., Hawkins, G. E., & Forstmann, B. U. (2016). Striatal activation reflects urgency in perceptual decision making. *NeuroImage*, *139*, 294–303. doi: 10.1016/j.neuroimage.2016.06.045

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, *27*(5), 1413–1432. doi: 10.1007/s11222-016-9696-4

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.

Waltz, J. A., Frank, M. J., Robinson, B. M., & Gold, J. M. (2007). Selective reinforcement learning deficits in schizophrenia support predictions from computational models of striatal-cortical dysfunction. *Biological psychiatry*, *62*(7), 756–764. doi: 10.1016/j.biopsych.2006.09.042

Watabe-Uchida, M., Eshel, N., & Uchida, N. (2017). Neural circuitry of reward prediction error. *Annal Review of Neuroscience*, *40*, 373–394. doi: 10.1146/annurev-neuro-072116-031109

Watanabe, S. (2013). A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, *14*, 867–897.

Wiecki, T. V., & Frank, M. J. (2013). A computational model of inhibitory control in frontal cortex and basal ganglia. *Psychological Review*, *120*(2), 329–355. doi: 10.1037/a0031542

Wiering, M., & vanOtterlo, M. (2012). *Reinforcement learning: State-of-the-art.* Springer.

Wilke, M., Kagan, I., & Andersen, R. A. (2012). Functional imaging reveals rapid reorganization of cortical activity after parietal inactivation in monkeys. *PNAS*, *109*(21), 8274–8279. doi: 10.1073/pnas.1204789109

Worthy, D. A., Hawthorne, M. J., & Otto, A. R. (2013). Heterogeneity of strategy use in the Iowa gambling task: A comparison of win-stay/lose-shift and reinforcement learning models. *Psychonomic Bulletin & Review*, *20*(2), 364–371. doi: 10.3758/s13423-012-0324-9

Yechiam, E., & Busemeyer, J. R. (2005). Comparison of basic assumptions embedded in learning models for experience-based decision making. *Psychonomic Bulletin & Review*, *12*(3), 387–402. doi: 10.3758/BF03193783

Yechiam, E., & Hochman, G. (2013). Loss-aversion or loss-attention: The impact of losses on cognitive performance. *Cognitive Psychology*, *66*(2), 212–231. doi: 10.1016/j.cogpsych.2012.12.001

Yeung, N., & Sanfey, A. G. (2004). Independent coding of reward magnitude and valence in the human brain. *Journal of Neuroscience*, *24*(28), 6258–6264. doi: 10.1523/JNEUROSCI.4537-03.2004

Zaghloul, K. A., Blanco, J. A., Weidemann, C. T., McGill, K., Jaggi, J. L., Baltuch, G. H., & Kahana, M. J. (2009). Human substantia nigra neurons encode unexpected financial rewards. *Science*, *323*(5920), 1496–1499. doi: 10.1126/science.1167342

Zeelenberg, M. (1999). Anticipated regret, expected feedback and behavioral decision making. *Journal of behavioral decision making*, *12*(2), 93–106. doi: 10.1002/(SICI)1099-0771(199906)12:2⟨93::AID-BDM311⟩3.0.CO;2-S

Zhang, Y., Larcher, K., Misic, B., & Dagher, A. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, *20*(1), 45–57. doi: 10.1109/42.906424

Zhang, Y., Larcher, K., Misic, B., & Dagher, A. (2017). Anatomical and functional organization of the human substantia nigra and its connections. *eLife*, *6*(e26653), 1–23.

doi: 10.7554/eLife.26653

Appendix A

Appendix Manuscript I

A.1 Bayesian mixed model ANOVA

In this section, we detail the results of the two-step model-comparison approach that we used for the Bayesian Mixed Model ANOVA.

In a first step, to determine what would be the base model for the subsequent analyses, we compared two models: the first one (M0) included only participants as random effects and the second one (M1) included also experiment as fixed effect. In the case of accuracy, a model that does not include experiment as fixed effect was preferred (BF_{M0}/ BF_{M1}=3.6), indicating that mean accuracy was mostly stable across experiments. In the case of RTs, a model that includes experiment as fixed effect was preferred (BF_{M1}/ BF_{M0}=1.4e9), indicating that mean RTs differed across experiments.

In a second step, we tested different combinations of models in which we varied the possible interactions between experiment and experimental manipulations and the experimental manipulations themselves. In the case of accuracy, all models were tested against M0 of the previous step of the analyses, while, in the case of RTs, these were tested against M1 of the previous analyses. The results are summarized in the following Supplementary Tables 1 and 2.

Finally, the two models with highest BF were compared to each other, to provide a simple assessment of the evidence in favor of the best model is. There was substantial evidence for the winning model in the ANOVA of accuracy analyses, M3, compared to its

runner-up, M8 (BF_{M3}/ BF_{M8}=8.6). There was anecdotal evidence for the winning model in the ANOVA of RT analyses, M5, compared to its runner-up, M10 (BF_{M5}/ BF_{M10}=1.76).

Table A.1: Bayes Factors of the ANOVA of accuracy.

Model	Random effects	Experiment interactions	Fixed Effects	BF	log(BF)
M2	participant	None	Valence	1.11e	-1-2.19
M3	participant	None	Feedback	2.26e7	16.93*
M4	participant	None	valence+feedback	2.53e6	14.74
M5	participant	None	valence*feedback	5.25e5	13.17
M6	participant	Valence	None	9.76e-2	-2.33
M7	participant	Valence	Valence	1.09e-2	-4.51
M8	participant	Valence	Feedback	2.64e6	14.79
M9	participant	Valence	valence+feedback	2.97e5	12.60
M10	participant	Valence	valence*feedback	6.19e4	11.03
M11	participant	Feedback	None	9.69e-2	-2.33
M12	participant	Feedback	Valence	1.08e-2	-4.53
M13	participant	Feedback	Feedback	1.38e6	14.13
M14	participant	Feedback	valence+feedback	1.54e5	11.95
M15	participant	Feedback	valence*feedback	3.21e4	10.38
M16	participant	valence+feedback	None	9.58e-3	-4.65
M17	participant	valence+feedback	Valence	1.07e-3	-6.84
M18	participant	valence+feedback	Feedback	1.62e5	12.00
M19	participant	valence+feedback	valence+feedback	1.83e4	9.81
M20	participant	valence+feedback	valence*feedback	3.81e3	8.25

Note. The preferred model is marked with an asterisk.

Table A.2: Bayes Factors of the ANOVA of response times.

Model	Random effects	Experiment interactions	Fixed Effects	BF	log(BF)
M2	participant	None	experiment+valence	3.39e24	56.48
M3	participant	None	experiment+feedback	7.88e-1	-0.24
M4	participant	None	experiment+valence+feedback	8.35e24	57.38
M5	participant	None	experiment+valence*feedback	1.23e27	62.37*
M6	participant	Valence	experiment	6.24e-1	-0.47
M7	participant	Valence	experiment+valence	1.54e24	55.69
M8	participant	Valence	experiment+feedback	5.17e-1	-0.66
M9	participant	Valence	experiment+valence+feedback	4.05e24	56.66
M10	participant	Valence	experiment+valence*feedback	6.99e26	61.81
M11	participant	Feedback	experiment	6.92e-2	-2.67
M12	participant	Feedback	experiment+valence	3.98e23	54.34
M13	participant	Feedback	experiment+feedback	6.10e-2	-2.80
M14	participant	Feedback	experiment+valence+feedback	1.18e24	55.43
M15	participant	Feedback	experiment+valence*feedback	1.90e26	60.51
M16	participant	valence+feedback	experiment	4.42e-2	-3.12
M17	participant	valence+feedback	experiment+valence	1.86e23	53.58
M18	participant	valence+feedback	experiment+feedback	4.10e-2	-3.19
M19	participant	valence+feedback	experiment+valence+feedback	5.95e23	54.74
M20	participant	valence+feedback	experiment+valence*feedback	1.13e26	59.99

Note. The preferred model is marked with an asterisk.

A.2 Reinforcement learning model analyses

In this section, we report some details about the reinforcement learning modelling procedure.

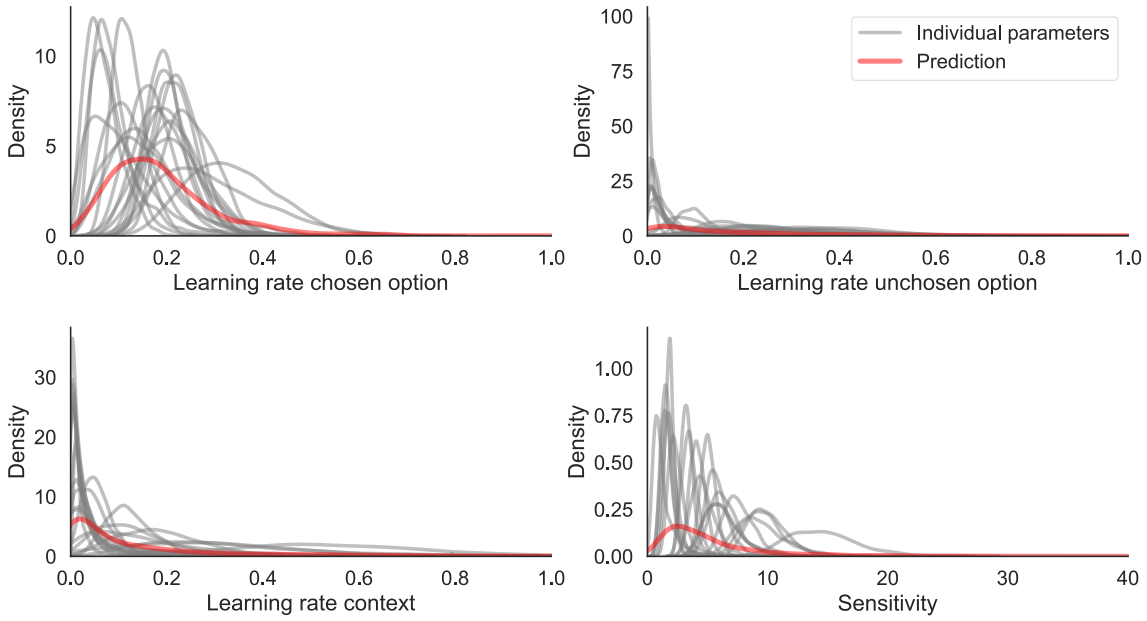


Figure A.1: *Posterior distributions of the hierarchical reinforcement learning (RL) model.* Posterior distributions of the RL model parameters at the individual level (grey lines) for the 20 participants of Experiment 1. Superimposed (red lines), are the distributions of the parameters used to generate predictions for new, non-observed participants (in Experiments 2, 3, and 4).

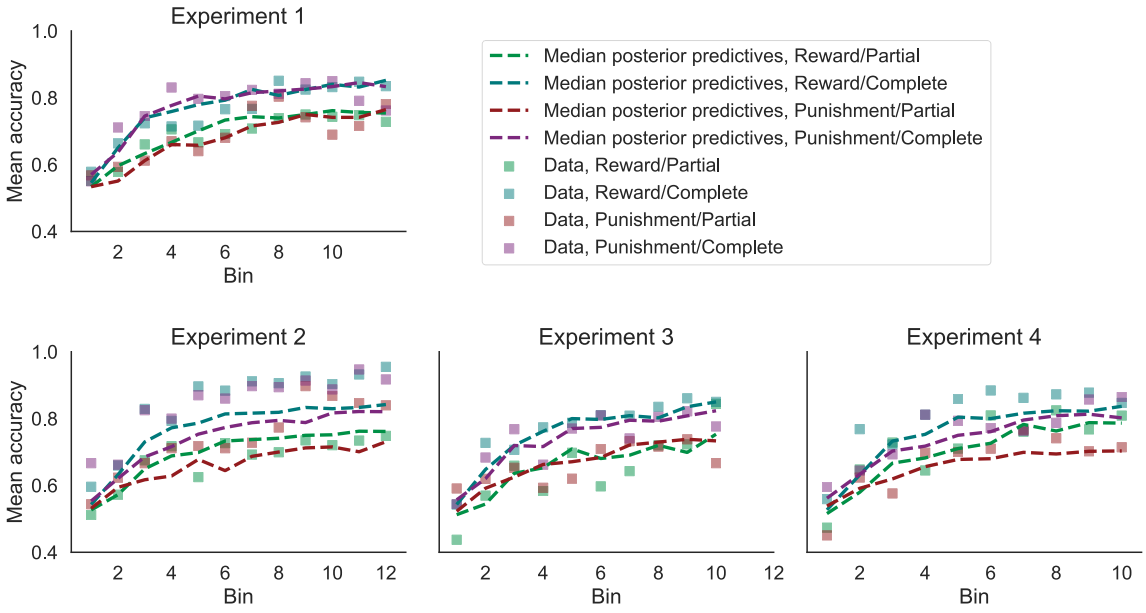


Figure A.2: *Posterior predictives of the hierarchical reinforcement learning (RL) model.* Posterior predictives for mean accuracy in binned trials, separately for learning contexts. Each bin corresponds to 12 trials, which means 3 trials per choice context. Mean accuracy was calculated separately across experiments, contexts, and bins. The dotted lines represent the median of the posterior predictive distribution of mean accuracy. Note that these are in-sample predictions for Experiment 1 (first row), and out-of-sample predictions for Experiments 2, 3, and 4 (second row).

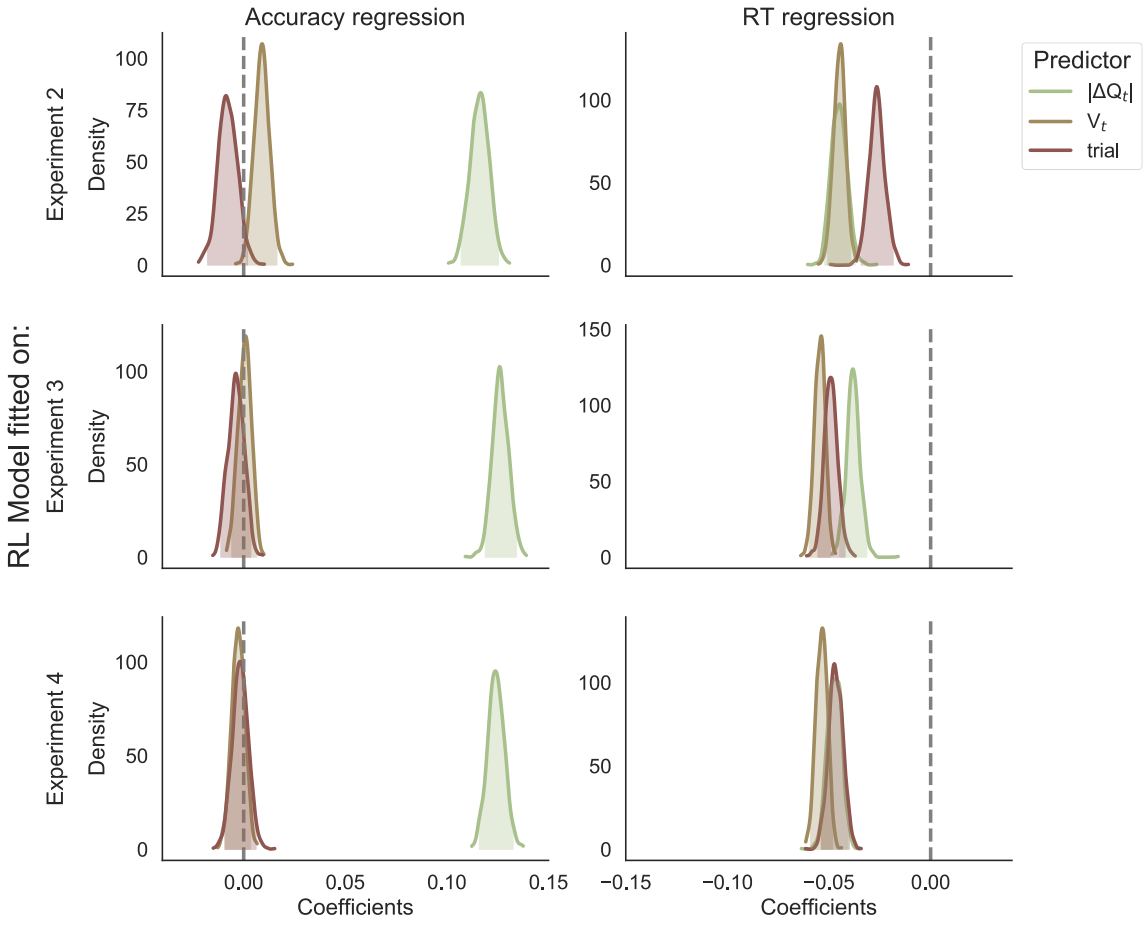


Figure A.3: *Control analyses for the reinforcement learning (RL) model.* Estimated posterior distributions of the linear model coefficients corresponding to the —Q—, V, and number of trial predictors. Separate models were tested to predict either accuracy (left column) or RTs (right column). Analyses were repeated by fitting the RL model on Experiment 2, 3, and 4 (top, middle, and bottom rows, respectively) and generating predictions for the remaining experiments (e.g., if the RL model was fitted on Experiment 2, it was then tested on Experiments 1, 3, and 4).

Table A.3: Bayes Factors for the linear model of accuracy.

Model	Random effects	Fixed Effects	BF	log(BF)
M1 (intercept model)	participant + experiment	trial	1	0
M2	participant + experiment	$\|\Delta Q_t\|$ + trial	7.6e183	423.4*
M3	participant + experiment	V_t + trial	1.3e5	11.8
M4	participant + experiment	$\|\Delta Q_t\|$ + V_t + trial	3.1e182	420.2

Note. The intercept model is M1. The winning model is indicated with a star.

Table A.4: Bayes Factors for the linear model of response times.

Model	Random effects	Fixed Effects	BF	log(BF)
M1 (intercept model)	participant + experiment	trial	1	0
M2	participant + experiment	$\|\Delta Q_t\|$ + trial	4.44e50	116.6
M3	participant + experiment	V_t + trial	2.80e107	247.4
M4	participant + experiment	$\|\Delta Q_t\|$ + V_t + trial	3.18e132	305.1*

Note. The intercept model is M1. The winning model is indicated with a star.

A.3 Diffusion decision model analyses

In this section, we report some details about the diffusion decision model analyses.

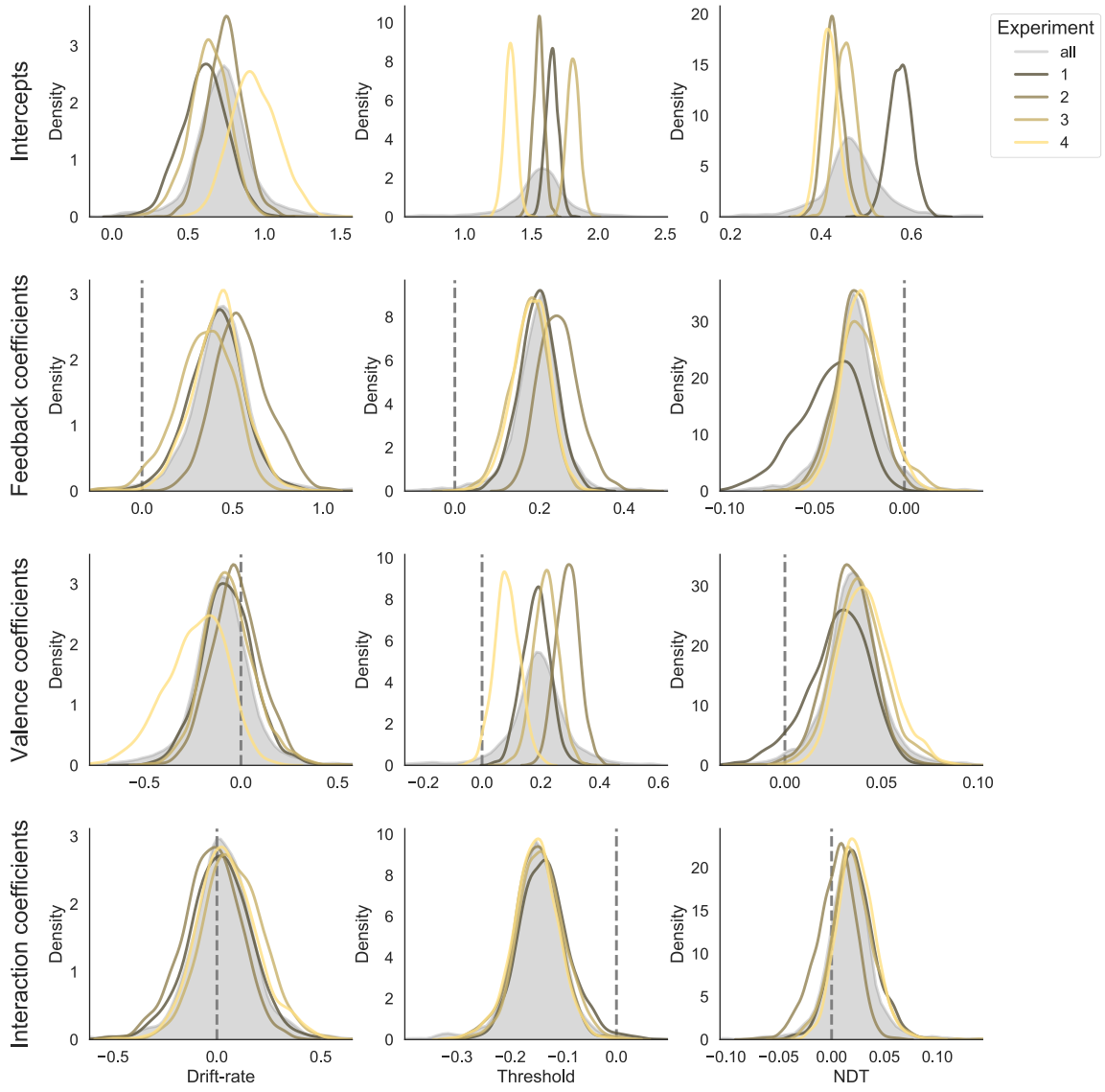


Figure A.4: *Posterior distributions of the group parameters of the hierarchical diffusion decision model.* Posterior distributions of the DDM parameters across experiments (grey areas) and within experiments. Because valence was coded as 0=reward/1=punishment, and feedback was coded as 0=partial/1=complete, and the interaction was the product of the two, intercepts (first row) correspond to the parameters in the reward-partial condition. Note that while the intercept of the drift-rates are identical across experiments, the thresholds appear to significantly differ, with the threshold being seemingly lower in Experiment 4 than in the other experiments. This is consistent with the fact that a shorter decision time window was allowed in Experiment 4 (see Table 2.1), which led to a reduction in the cautiousness in participants responses. Beyond this effect of experiments on DDM parameters intercepts, note that the pattern of the coefficients indexing the effects of the experimental factors (Feedback, valence and interactions) on DDM parameters was very consistent across experiments.

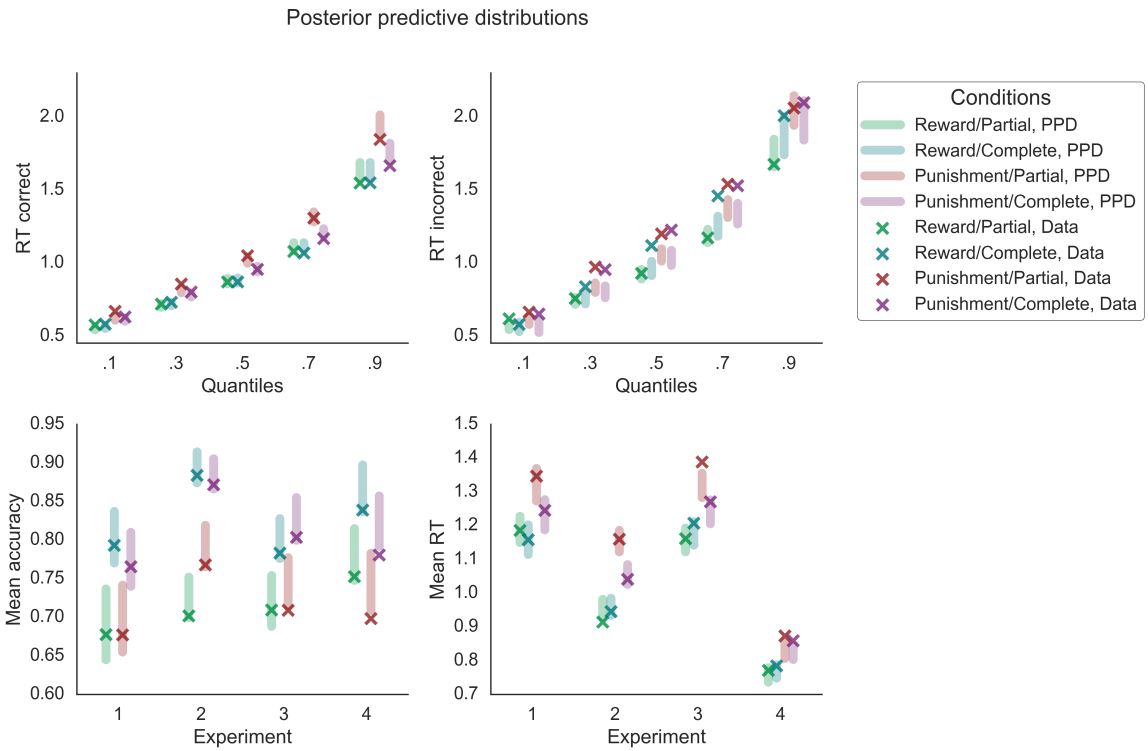


Figure A.5: *Posterior predictives of the hierarchical diffusion decision model.* Posterior predictive distributions for mean accuracy (bottom-left), mean RT (bottom-right), RT quantiles of correct (top-left) and incorrect (top-right) responses. To assess how well the model fits the observed behavioral patterns, these measures were separately calculated across experiments and experimental conditions. The shaded areas represent the 95% Bayesian Credible Intervals, while the crosses represent the summary of the data.

A.4 Diffusion decision model parameter recovery

We performed parameter recovery of the Bayesian hierarchical diffusion decision model(DDM) used in the main analyses of this study. We generated data for four experiments using a simple DDM (with no across-trial variability), with the same number of participants and trials as in our study.

The generating group parameters (Table A.5) were selected in order to generate a similar performance to the one observed across the experiments (Figure A.6). Participants parameters were sampled from the group distributions and NDT and threshold intercepts were lowered in Experiment 4 of .3 and .05, respectively.

We fitted the DDM following the same procedure used to fit the real data collected in the four experiments, as described in Section 2.2. To assess the quality of parameter recovery, we plotted the generating parameter values against a summary (mean and mode) of the estimated posterior distributions of the 89 participants (Figure A.7). In general, all parameters were well recovered, although for some parameters we observe a shrinkage towards the group mean (e.g., for the interaction coefficient of the drift-rate and for the valence coefficient of the threshold) which is a typical feature of hierarchical models.

Table A.5: Generating parameters.

	Drift-rate	Threshold	NDT
Mean intercept	0.6	1.5	0.3
SD intercept	0.2	0.1	0.1
Mean coefficients (valence, feedback, interaction)	0.00,0.50, 0.00	0.20, 0.10, -0.10	0.12, 0.00, 0.10
SD coefficients (valence, feedback, interaction)	0.10, 0.05, 0.10	0.05, 0.05, 0.03	0.04, 0.10, 0.08

Note. The generating parameters at the dataset level were used for the parameter recovery.

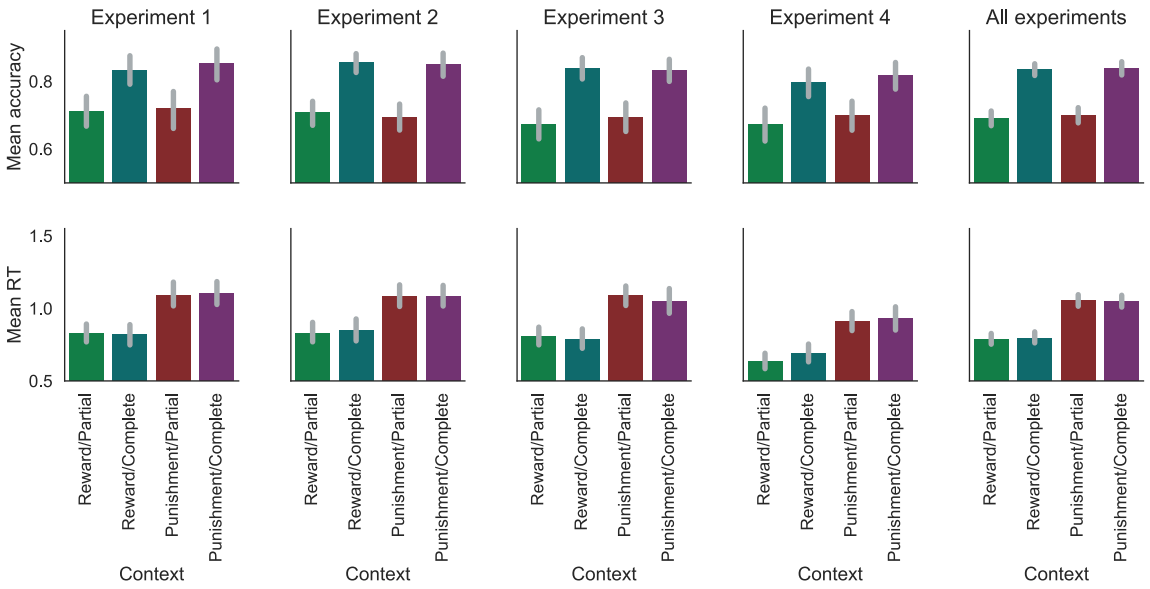


Figure A.6: *Simulated data.* Mean accuracy and response times (RTs) of the simulated data, separately by experiment and by context.

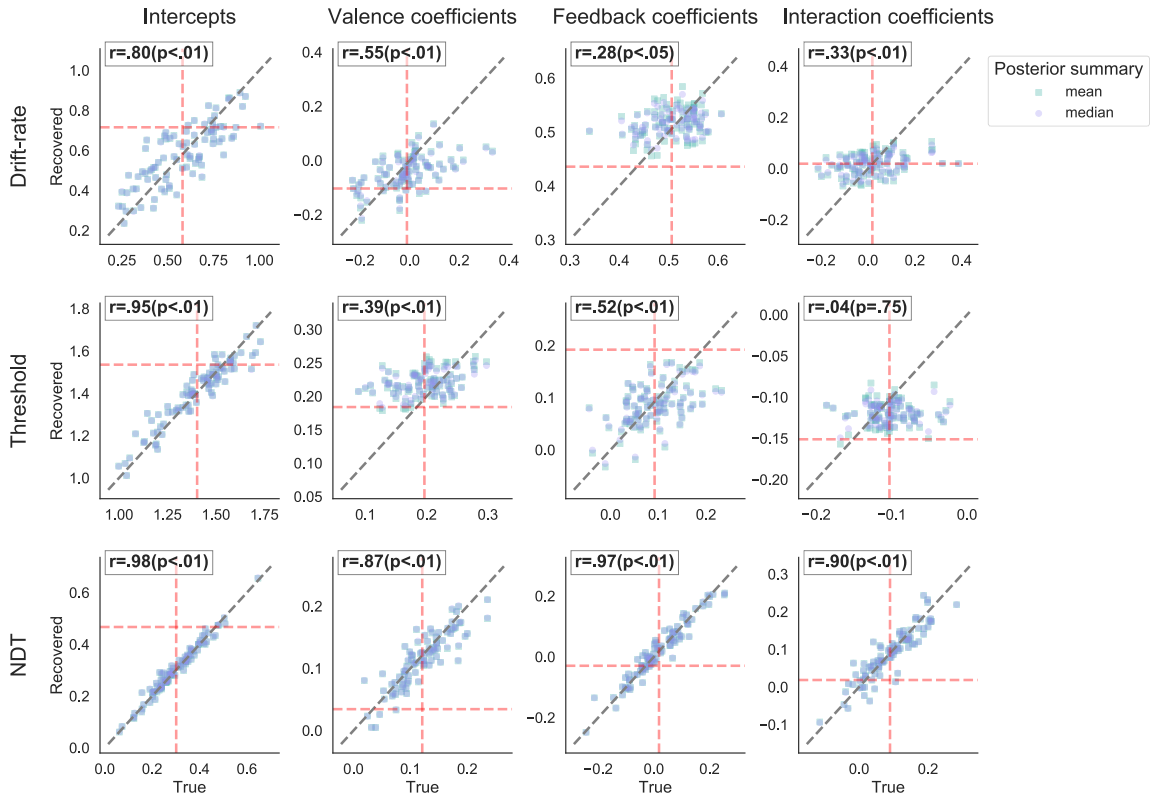


Figure A.7: *True against recovered diffusion decision model individual parameters.* The dotted grey lines represent the identity lines, while the red dotted lines are the group mean parameters. We also calculated correlations between the true and the mean recovered individual parameters, indicated by the Pearson's statistics. Apart from the threshold interaction coefficient parameter, all correlations are significant. We also plotted the generating group means against the mean of the posterior distribution.

Appendix B

Appendix Manuscript II

B.1 Bayesian hierarchical regression models

A graphical representation of the linear regression model of accuracy can be seen in Figure B.1 on the right. The model, model priors, and hyper-priors were specified as:

$$\mu_\alpha \sim \mathcal{N}(0,5); \sigma_\alpha \sim \mathcal{HN}(0,5)$$

$$\boldsymbol{\mu_\beta} \sim \mathcal{N}(0,5); \boldsymbol{\sigma_\beta} \sim \mathcal{HN}(0,5)$$

$$\alpha \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha); \beta \sim \mathcal{N}(\mu_\beta, \sigma_\beta)$$

$$p_{t,s} = logit(\mathbf{X_{t,s}}\boldsymbol{\beta_s} + \alpha_s); \mathrm{acc}_{t,s} \sim \mathrm{Bern}(p_{t,s})$$

where α and $\boldsymbol{\beta}$ are, respectively, the intercept and the vector of coefficients, and $\mathbf{X}$ is the predictors matrix. $\mathcal{N}$ is a normal distribution with parameters mean and standard-deviation, $\mathcal{HN}$ is a half-normal distribution with parameters mean and standard-deviation, and Bern is a Bernoulli distribution with a success probability parameter.

A graphical representation of the linear regression model of RT can be seen in Figure B.1 on the left. The model, model priors, and hyper-priors were specified as:

$$\mu_\alpha \sim \mathcal{N}(0,5); \sigma_\alpha \sim \mathcal{HN}(0,5)$$

$$\boldsymbol{\mu_\beta} \sim \mathcal{N}(0,5); \boldsymbol{\sigma_\beta} \sim \mathcal{HN}(0,5)$$

$$\sigma \sim \mathcal{HN}(0,5)$$

$$\alpha \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha); \beta \sim \mathcal{N}(\mu_\beta, \sigma_\beta)$$

$$\hat{\mathrm{RT}}_{t,s} = \mathbf{X_{t,s}}\boldsymbol{\beta_s} + \alpha_s; \mathrm{RT}_{t,s} \sim \mathcal{N}(\hat{\mathrm{RT}}_{t,s}, \sigma)$$

where α and $\boldsymbol{\beta}$ are, respectively, the intercept and the vector of coefficients, σ is the noise term, and $\mathbf{X}$ is the predictors matrix. $\mathcal{N}$ is a normal distribution with parameters mean and standard-deviation, and $\mathcal{HN}$ is a half-normal distribution with parameters mean and standard-deviation.

Models specification followed the approach of (Gelman et al., 2014), while priors followed the prior choice recommendations in the Stan manual.

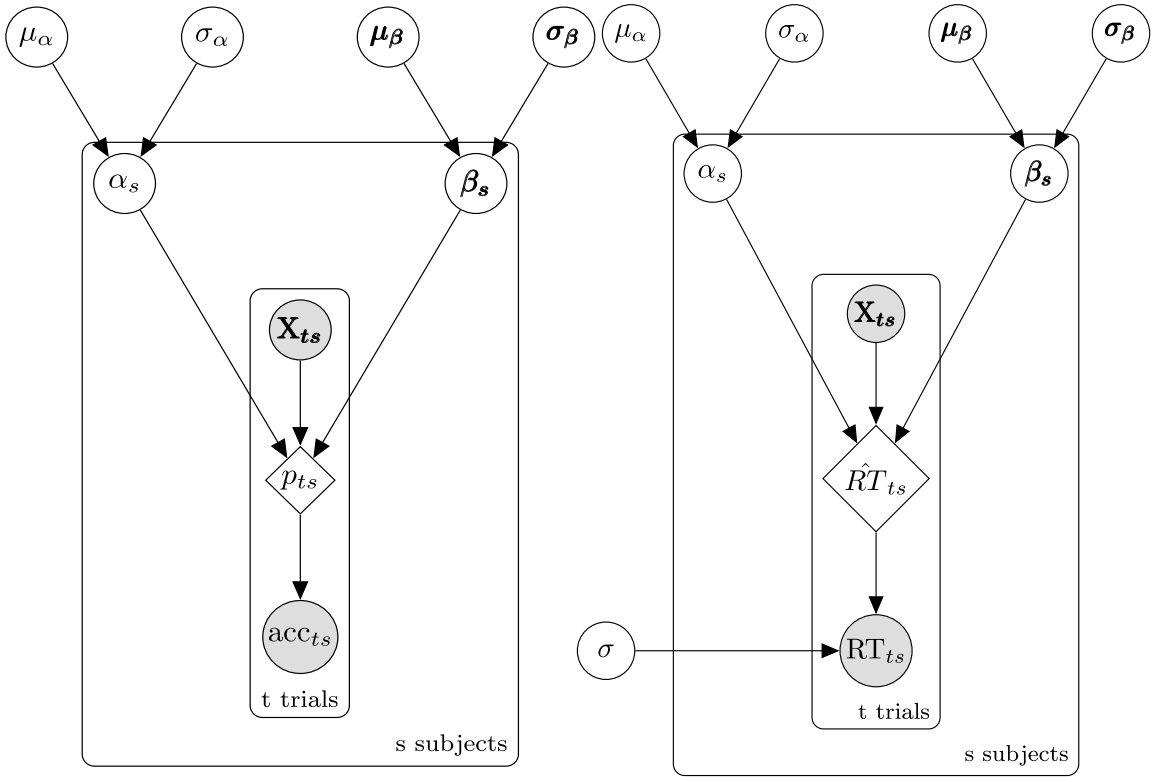


Figure B.1: Graphical representation of the Bayesian hierarchical linear regression (on the left) and logistic regression (on the right) models. Shaded nodes represent observed variables, while non-shaded nodes represent latent variables. Squared nodes represent deterministic variables.

B.2 Bayesian hierarchical cognitive models

A graphical representation of the full reinforcement learning model (i.e., with separate learning rates η for positive and negative prediction error and increasing sensitivity θ) can be seen in Figure B.2. The model, model priors, and hyper-priors were specified as:

$$\mu_{\eta^{\pm}} \sim \mathcal{N}(0, .8); \sigma_{\eta^{\pm}} \sim \mathcal{HN}(0, .5)$$

$$\eta^{\pm}\pm \sim \phi(\mathcal{N}(\mu_{\eta^{\pm}}, \sigma_{\eta^{\pm}}))$$

$$\mu_b \sim \mathcal{N}(5, .5); \sigma_b \sim \mathcal{HN}(0, .5)$$

$$b \sim exp(\mathcal{N}(\mu_b, \sigma_b))$$

$$\mu_c \sim \mathcal{N}(-.5, .5); \sigma_c \sim \mathcal{HN}(0, .5)$$

$$c \sim exp(\mathcal{N}(\mu_c, \sigma_c))$$

$$\text{acc}_{t,s} \sim \text{Bern}(logit(p_{t,s}))$$

where $\eta^{\pm}$ are the learning rates for positive and negative prediction errors, b and c are the parameters that define the increase of θ in time (see Equation 3.3). $\mathcal{N}$ is a normal distribution with parameters mean and standard-deviation, $\mathcal{HN}$ is a half-normal distribution with parameters mean and standard-deviation, and Bern is a Bernoulli distribution with a success probability parameter.

A graphical representation of the Bayesian hierarchical diffusion decision model (DDM) can be seen in Figure B.3. The model, model priors, and hyper-priors were specified as:

$$\mu_v \sim \mathcal{N}(1, 2); \sigma_v \sim \mathcal{HN}(0, 2)$$

$$v \sim \mathcal{N}(\mu_v, \sigma_v)$$

$$\mu_a \sim \Gamma(1, 2); \sigma_a \sim \mathcal{HN}(0, .5)$$

$$a \sim \mathcal{N}(\mu_a, \sigma_a)$$

$$\mu_{T_{er}} \sim \mathcal{U}(0, 1); \sigma_{T_{er}} \sim \mathcal{HN}(0, .5)$$

$$T_{er} \sim \mathcal{N}(\mu_{T_{er}}, \sigma_{T_{er}})$$

$$\hat{v}_{t,s} = \begin{cases} v_s, & \text{if } \text{acc}_{t,s} = 1 \\ -v_s, & \text{if } \text{acc}_{t,s} = -1 \end{cases}$$

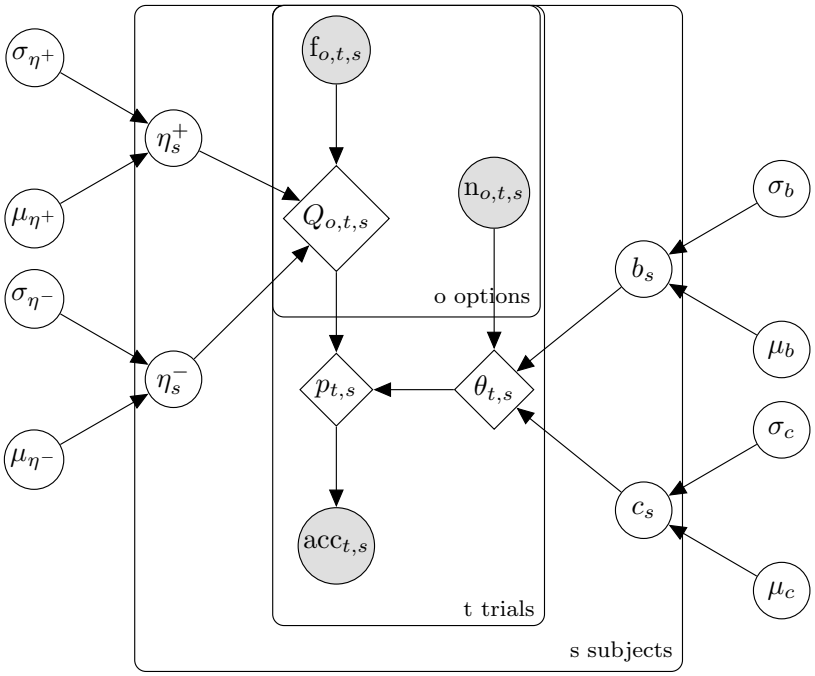


Figure B.2: Graphical representation of the Bayesian hierarchical full reinforcement learning (RL) model. Shaded nodes represent observed variables, while non-shaded nodes represent latent variables. Squared nodes represent deterministic variables.

$$\mathrm{RT}_{t,s} \sim Wiener(\hat{v}_{t,s}, a_s, T_{er_s}, .5)$$

where $Wiener$ is the Wiener distribution (Navarro & G.Fuss, 2009) with parameters drift rate, threshold, non-decision time, and relative starting-point, Γ is a gamma distribution with parameters shape and scale, and $\mathcal{U}$ is a uniform distribution with parameters lower and upper bounds. In our case, the starting-point is fixed at .5, since the options were randomly positioned on the right or left of a fixation cross across the trials and we therefore assumed that participants were not biased towards a particular position.

A graphical representation of the Bayesian hierarchical full reinforcement learning diffusion decision model (RLDDM) can be seen in Figure B.4. The model priors and hyper-priors were specified as:

$$\mu_{v_{\max}} \sim \mathcal{N}(0,1); \sigma_{v_{\max}} \sim \mathcal{HN}(0,.5)$$

$$v_{\max} \sim exp(\mathcal{N}(\mu_{v_{\max}}, \sigma_{v_{\max}})$$

$$\mu_{v_{\mathrm{mod}}} \sim \mathcal{N}(-1,1); \sigma_{v_{\mathrm{mod}}} \sim \mathcal{HN}(0,.5)$$

$$v_{\mathrm{mod}} \sim exp(\mathcal{N}(\mu_{v_{\mathrm{mod}}}, \sigma_{v_{\mathrm{mod}}}))$$

$$\mu_{a_{\mathrm{fixed}}} \sim \Gamma(1,2); \sigma_{a_{\mathrm{fixed}}} \sim \mathcal{HN}(0,.5)$$

$$a_{\mathrm{fixed}} \sim \mathcal{N}(\mu_{a_{\mathrm{fixed}}}, \sigma_{a_{\mathrm{fixed}}})$$

$$\mu_{a_{\mathrm{mod}}} \sim \mathcal{N}(0,.1); \sigma_{a_{\mathrm{mod}}} \sim \mathcal{HN}(0,.1)$$

$$a_{\mathrm{mod}} \sim \mathcal{N}(\mu_{a_{\mathrm{mod}}}, \sigma_{a_{\mathrm{mod}}})$$

Priors and hyper-priors for $\eta^{\pm}$ were the same as in the RL model, while for T_{er} they were the same as in the DDM.

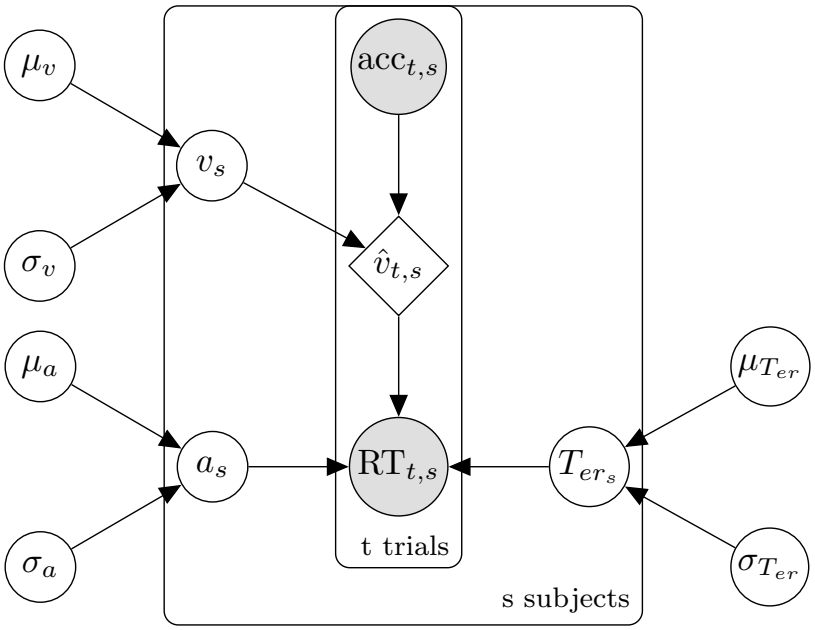


Figure B.3: Graphical representation of the Bayesian hierarchical diffusion decision model (DDM). Shaded nodes represent observed variables, while non-shaded nodes represent latent variables. Squared nodes represent deterministic variables.

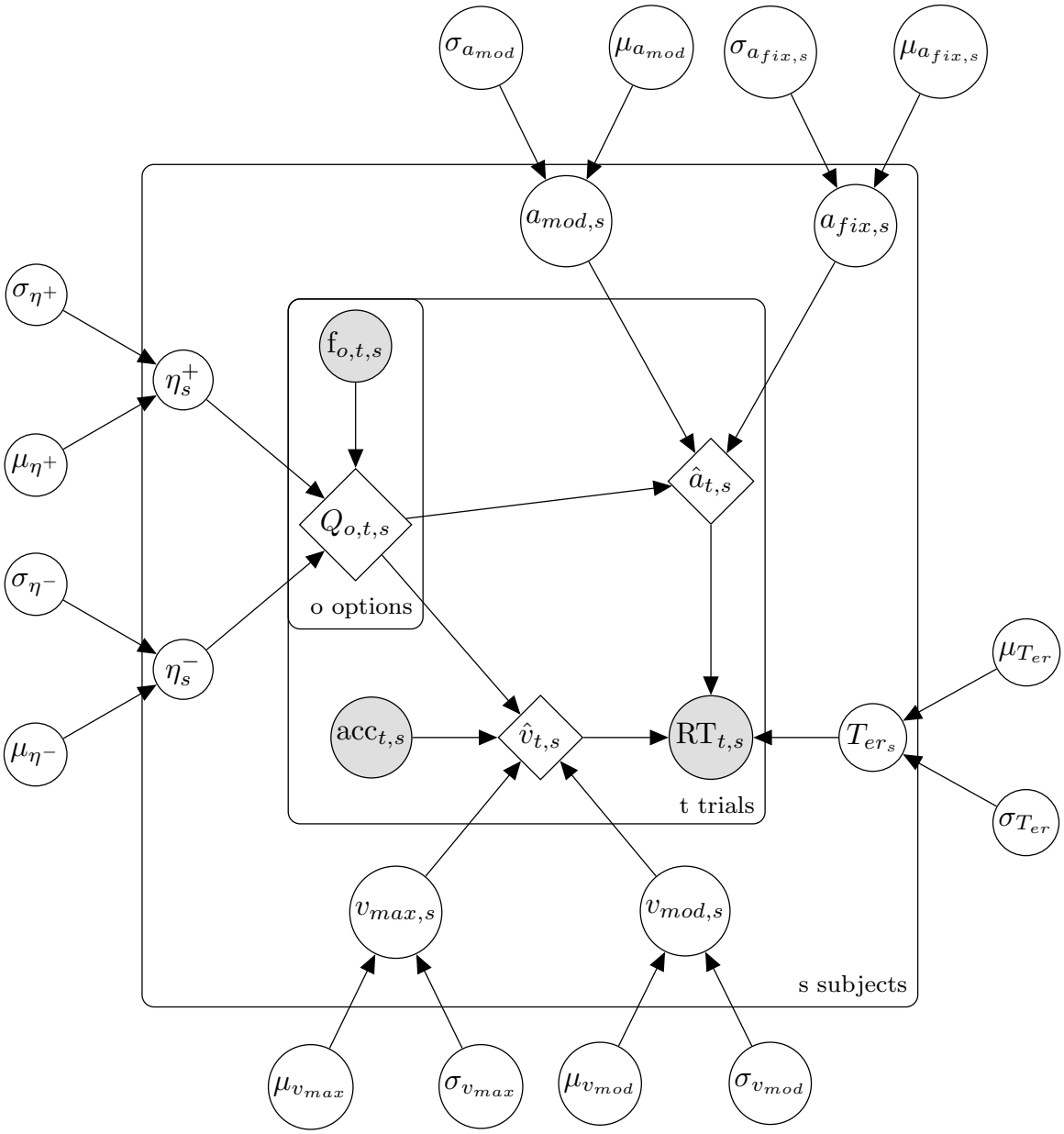


Figure B.4: Graphical representation of the Bayesian hierarchical full reinforcement learning diffusion decision model (RLDDM). Shaded nodes represent observed variables, while non-shaded nodes represent latent variables. Squared nodes represent deterministic variables. Vectors are represented in bold font.

Table B.1: Group parameter estimates of the full reinforcement learning model.

Parameter	M	SD	2.5% percentile	97.5% percentile
$\phi(\mu_{\eta^+})$	0.07	0.02	0.04	0.10
σ_{η^+}	0.57	0.11	0.38	0.82
$\phi(\mu_{\eta^-})$	0.24	0.04	0.17	0.34
σ_{η^-}	0.32	0.21	0.02	0.79
$exp(\mu_\theta)$	0.35	0.05	0.26	0.48
σ_θ	0.63	0.12	0.44	0.88

Note. The best fitting reinforcement model had separate learning rates η^+ and η^- for positive and negative prediction errors, and fixed sensitivity θ throughout learning. Note that μ_{η^+}, μ_{η^-}, and μ_θ were transformed for interpretability.

B.3 Results of parameter estimation

In this section we report the results of estimation of the group parameters for the best fitting models in each class of models (i.e., RL, DDM, RLDDM, and the models of Pedersen et al., 2017).

Table B.2: Group parameter estimates of the diffusion decision model.

Parameter	M	SD	2.5% percentile	97.5% percentile
$\mu_{v,\text{diff}}$	0.67	0.07	0.54	0.81
$\sigma_{v,\text{diff}}$	0.33	0.05	0.24	0.45
$\mu_{v,\text{easy}}$	1.28	0.08	1.11	1.44
$\sigma_{v,\text{easy}}$	0.42	0.07	0.31	0.58
$\mu_{a,\text{AB}}$	1.97	0.06	1.86	2.09
$\sigma_{a,\text{AB}}$	0.27	0.05	0.19	0.37
$\mu_{a,\text{AC}}$	2.11	0.08	1.94	2.27
$\sigma_{a,\text{AC}}$	0.39	0.07	0.28	0.54
$\mu_{a,\text{BD}}$	1.81	0.06	1.69	1.94
$\sigma_{a,\text{BD}}$	0.31	0.05	0.22	0.42
$\mu_{a,\text{CD}}$	1.70	0.05	1.61	1.80
$\sigma_{a,\text{CD}}$	0.23	0.04	0.17	0.32
$\mu_{T_{er}}$	0.75	0.02	0.71	0.80
$\sigma_{T_{er}}$	0.13	0.02	0.10	0.17

Note. The diffusion decision model had separate drift rate v for easy (between AC and BD) and difficult (between AB and CD) choices, a different threshold a for each pair of options, and one non-decision time T_{er}.

Table B.3: Group parameter estimates of the full reinforcement learning diffusion decision model.

Parameter	M	SD	2.5% percentile	97.5% percentile
$\phi(\mu_{\eta^+})$	0.07	0.02	0.03	0.12
σ_{η^+}	0.75	0.15	0.50	1.09
$\phi(\mu_{\eta^-})$	0.08	0.02	0.05	0.14
σ_{η^-}	0.58	0.13	0.37	0.87
$exp(\mu_{v_{\text{mod}}})$	0.48	0.10	0.32	0.70
$\sigma_{v_{\text{mod}}}$	0.85	0.14	0.59	1.14
$exp(\mu_{v_{\text{max}}})$	3.47	0.25	2.98	3.98
$\sigma_{v_{\text{max}}}$	0.31	0.07	0.20	0.47
$\mu_{a_{\text{fixed}}}$	1.00	0.20	0.62	1.39
$\sigma_{a_{\text{fixed}}}$	0.97	0.14	0.73	1.26
$\mu_{a_{\text{mod}}}$	-0.010	0.006	-0.021	0.001
$\sigma_{a_{\text{mod}}}$	0.027	0.004	0.020	0.037
$\mu_{T_{er}}$	0.76	0.03	0.71	0.81
$\sigma_{T_{er}}$	0.13	0.02	0.10	0.17

Note. The full reinforcement model had separate learning rates η^+ and η^- for positive and negative prediction errors, two parameters to describe the non-linear mapping between the difference in values and the drift rate, a scaling parameter v_{mod}, and an asymptote v_{max}, one fixed threshold parameter a_{fixed}, one value-modulation parameter a_{mod}, and finally one non-decision time T_{er}. Note that μ_{η^+}, μ_{η^-}, $\mu_{v_{\text{mod}}}$, and $\mu_{v_{\text{max}}}$ were transformed for interpretability.

Table B.4: Group parameter estimates of the model of Pedersen et al. (2017).

Parameter	M	SD	2.5% percentile	97.5% percentile
$\phi(\mu_{\eta^+})$	0.09	0.02	0.05	0.14
σ_{η^+}	0.57	0.11	0.39	0.81
$\phi(\mu_{\eta^-})$	0.17	0.03	0.12	0.25
σ_{η^-}	0.47	0.13	0.24	0.75
$exp(\mu_m)$	0.14	0.01	0.12	0.17
σ_m	0.41	0.08	0.27	0.60
$exp(\mu_{bp})$	0.013	0.005	0.005	0.023
σ_{bp}	1.321	0.237	0.894	1.821
μ_{bb}	1.85	0.06	1.74	1.95
σ_{bb}	0.27	0.04	0.20	0.36
$\mu_{T_{er}}$	0.75	0.03	0.70	0.80
$\sigma_{T_{er}}$	0.13	0.02	0.10	0.17

Note. The model of Pedersen et al. (2017) had separate learning rates η^+ and η^- for positive and negative prediction errors, a drift rate scaling parameter m, two parameters to define the trial-by-trial decrease of threshold, bp and bb, and one non-decision time T_{er}. Note that μ_{η^+}, μ_{η^-}, μ_m, and μ_{bp} were transformed for interpretability.

B.4 Parameter recovery

Models in decision making, and particularly RL models, can suffer from poor identifiability and parameter recovery, especially when the information content in the data is low (Spektor & Kellen, 2018). To alleviate this concern, we conducted a parameter recovery study for the most complex model, the one that is most likely to suffer from poor identifiability. We ran 10 independent simulations based on the original dataset (i.e., with the same number of participants and trials per participant) and using the most likely samples of the estimated joint posterior distribution. We then estimated the group and individual parameters for each individual synthetic dataset as described in the Methods section.

We inspected the results in three different ways: (1) correlations across the samples of the group parameters (i.e., mean and standard deviations); (2) ability to correctly infer the group means; (3) ability to correctly infer the individual parameters, summarized as mean and median of the individual posterior distributions.

The correlations across samples can be seen in Figure B.5, obtained by averaging the correlation matrices across the independent simulations, together with the SD across correlations. The highest observed significant positive correlation (ρ=.19, p¡.05) was found between $\sigma_{v_{\max}}$ and $\sigma_{v_{\mathrm{mod}}}$, while the highest significant negative correlation (ρ=-.15, p¡.05) was found between $\mu_{v_{\max}}$ and $\mu_{v_{\mathrm{mod}}}$.

The posterior distributions of the mean group parameters for the 10 independent simulations can be seen in Figure B.6. While most group means and standard deviations were successfully recovered, $\sigma_{v_{\max}}$ was sometimes underestimated.

Finally, recovery of individual-level parameters is shown in Figure B.7. Most parameters recovered well and showed a slight shrinkage to the mean (which is a feature of hierarchical models).

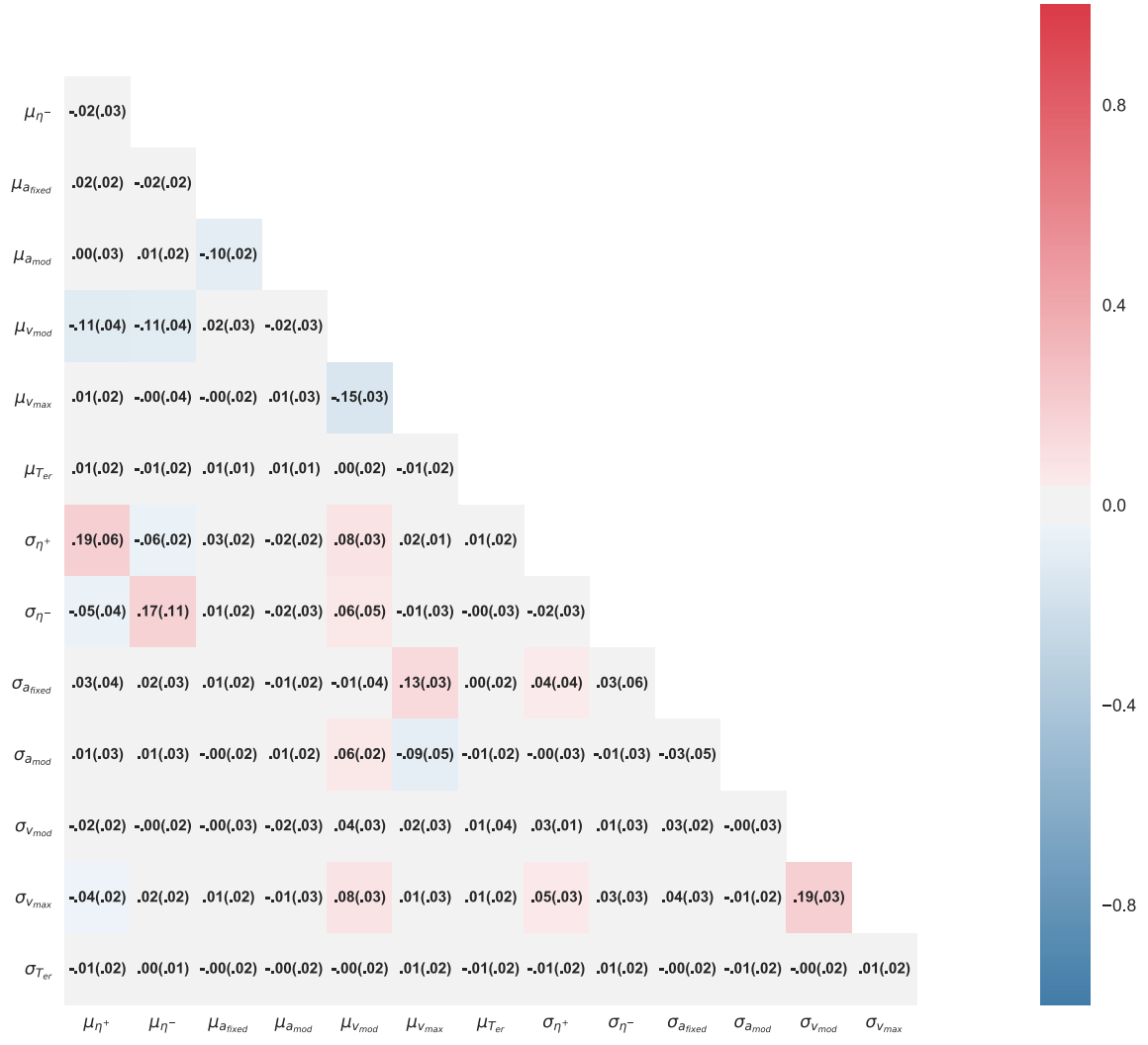


Figure B.5: Mean (and, between brackets, SD) of the correlation between the samples of the group-level parameters, across simulations.

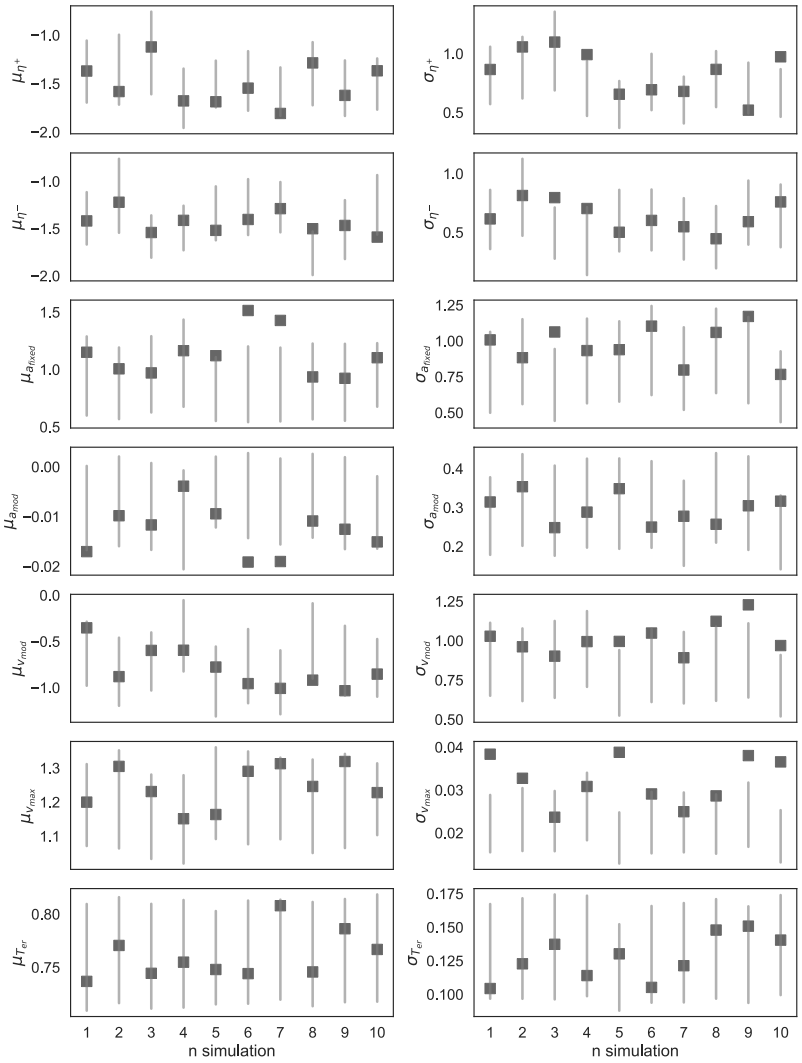


Figure B.6: 95% BCI of the estimated mean and standard deviation of the group parameter distributions (grey lines) and true generating group parameters (grey squared dots).

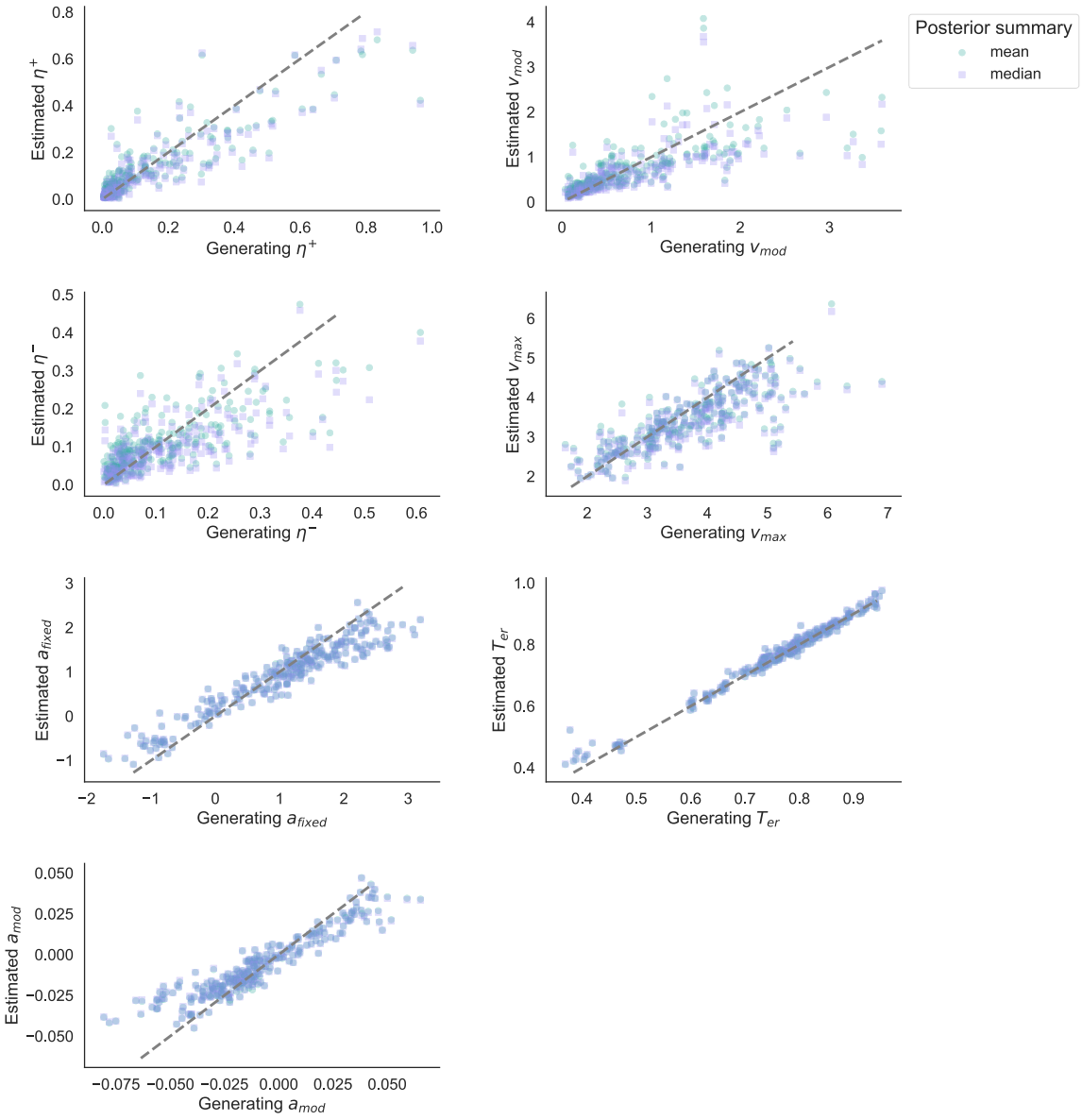


Figure B.7: Scatter plot of the estimated individual parameters, summarized as mean and median of the posterior distribution, against the true generating parameters. All 10 simulations are plotted here together.

Appendix C

Appendix Manuscript III

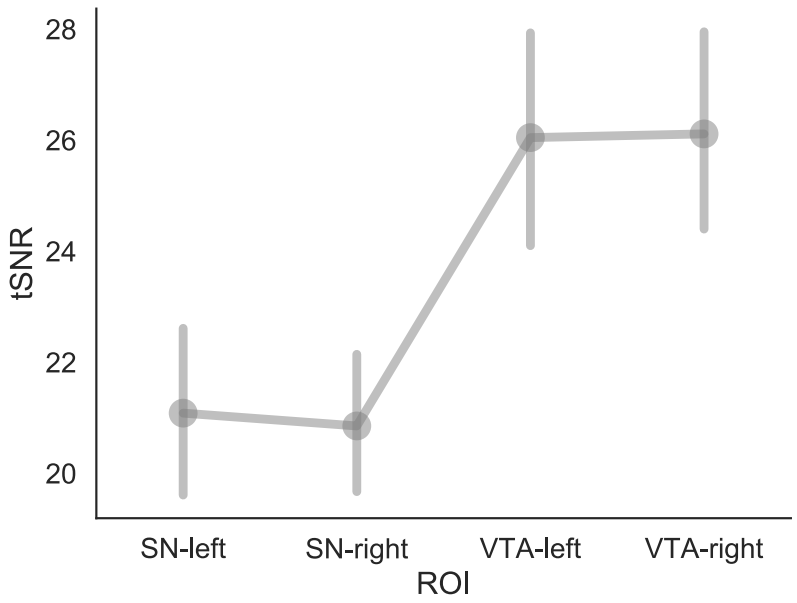


Figure C.1: Temporal signal-to-noise ratio (tSNR) across regions of interests (ROI): left and right substantia nigra (SN) and left and right ventral tegmental area (VTA). Bars represent 95% confidence intervals.

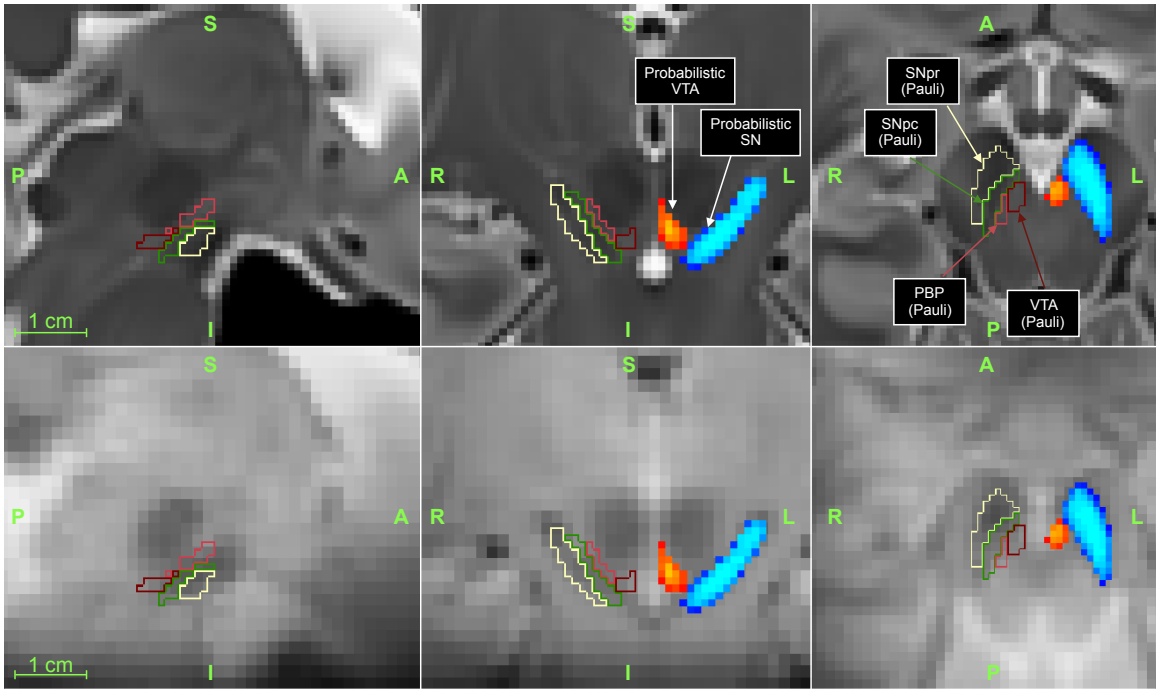


Figure C.2: Hard lines: Pauli et al. (2018)'s deterministic masks for the ventral tegmental area (VTA), parabrachial pigmented nucleus (PBP) – which together form what is usually referred to as VTA – substantia nigra pars reticulata (SNpr), and and pars compacta(SNpc) – which together form the SN. Red to yellow gradients: probabilistic map of the VTA, estimated in the current study. Blue to light-blue gradient: probabilistic map of the SN, estimated in the current study. Note that, in the probabilistic maps, darker colors represent lower probability for a voxel to belong to the ROI. Only voxels with probability higher than 30% were kept. In the top row, the background image is Pauli et al. (2018)'s template, while the mean functional image of the current study, in MNI space, is the background in the bottom row.

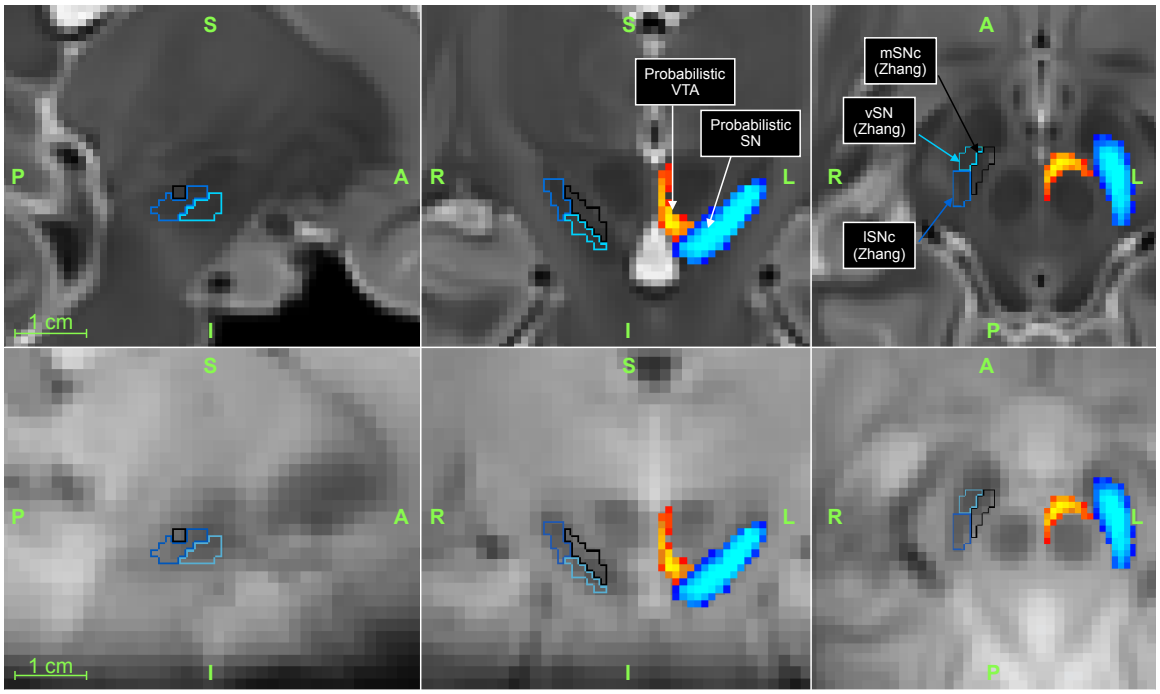


Figure C.3: Hard lines: Zhang et al. (2017)'s deterministic masks for the medial and lateral parts of the substantia nigra pars compacta (mSNc and lSNc) and for the ventral part of the substantia nigra (vSN) – which together form the SN. Red to yellow gradients: probabilistic map of the VTA, estimated in the current study. Blue to light-blue gradient: probabilistic map of the SN, estimated in the current study. Note that, in the probabilistic maps, darker colors represent lower probability for a voxel to belong to the ROI. Only voxels with probability higher than 30% were kept. In the top row, the background image is Pauli et al. (2018)'s template, while the mean functional image of the current study, in MNI space, is the background in the bottom row.

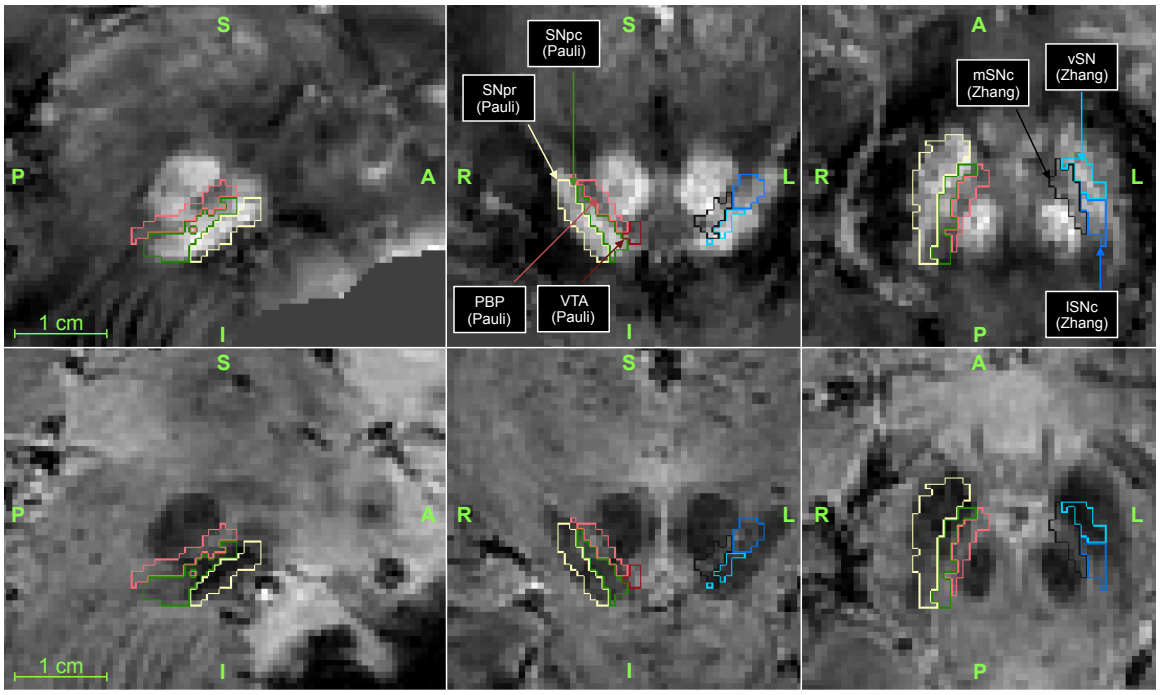


Figure C.4: Zhang et al. (2017)'s and Pauli et al. (2018)'s deterministic masks in one participant's individual space. The image on the top row is the quantitative susceptibility mapping (QSM) and the image on the bottom is the mean of the third and fourth echo of the T_2^*-weighted image.

Appendix D

Curriculum Vitae

Curriculum Vitae

LAURA FONTANESI, PHD

Postdoctoral researcher in Economic Psychology (Jörg Rieskamp's lab). Research on cognitive and neural mechanisms of reward-based decision and learning processes.

EDUCATION

- **PhD student in Economic Psychology (October 2014 – January 2019),** University of Basel, Basel, Switzerland; Research on cognitive and neural mechanisms of reward-based decision and learning processes, under supervision of Prof. Dr. Jörg Rieskamp and Prof. Dr. Sebastian Gluth;
- **Visiting PhD student (February 2017 - October 2017)**, University of Amsterdam, Amsterdam, The Netherlands. Conducting a research in Prof. Dr. Birte Forstmann group at the University of Amsterdam, co-supervised by Prof. Joerg Rieskamp and Prof. Sebastian Gluth from University of Basel. This research, funded by the Swiss National Science Foundation, focuses on the role of dopaminergic nuclei in economic decision-making, using ultra-high resolution 7Tesla magnetic resonance imaging;
- **Research Master's Degree in Psychology, major in Psychological Methods, minor in Brain and Cognition (September 2012 – August 2014)**, University of Amsterdam, Amsterdam, The Netherlands; Thesis Title: "*Evidence Accumulation in an Expanded Judgment Task: a 'Model-In-The-Middle' Approach*"; Thesis Supervisors: Prof. Dr. Birte Forstmann and Dr. Leendert Van Maanen;
- **Bachelor's Degree in Science and Technology of Cognitive Psychology (September 2008 – December 2011)**, 110 cum laude/110, University of Trento, Trento, Italy; Thesis Title: "*Who is afraid of statistics? A survey among the students of Cognitive Science*"; Thesis Supervisor: Prof. Dr. Luisa Canal.

ACADEMIC EXPERIENCE

ONGOING PROJECTS

- "*EEG-Signature of Reward-Based Decision-Making in a Gambling Task: How does Learning Biases the Decision Process?*", in collaboration with Sebastian Gluth, and Jörg Rieskamp;
- "*The Neural Basis of Sequential Decision-Making: How the brain decides to stop foraging under risk.*", in collaboration with Sebastian Gluth and Amitai Shenhav.

ORGANIZATION

- Was part of the core organizational committee of the **2016 JDMx Meeting for Early-Career Researchers**. The JDMx meeting for early-career researchers is a platform for PhD students and early post docs active in the judgment and decision making community. This was the ninth annual edition of the meeting, taking place from June 8th until June 11th 2016 in Basel, Switzerland.

TEACHING

- **Teaching the Psychology Bachelor's course: "Statistics in R" (September 2018 – now)**, University of Basel, Basel, Switzerland;
- **Teaching the Psychology Bachelor's course: "Project seminar" (September 2018 – now)**, University of Basel, Basel, Switzerland;
- **Teaching Assistant for the Psychology Research Master course: "Programming Skills: R & Matlab" (January 2014 – February 2014)**, University of Amsterdam, Amsterdam, The Netherlands;
- **Teaching Assistant for the Psychology Bachelor course: "Psychometrics" (April 2011 – September 2011)**, University of Trento, Trento, Italy.

Research Internships

- **Research Internship in Cognitive Psychology (February 2013 – July 2013)**, University of Amsterdam, Amsterdam, The Netherlands; Internship Title: *"Social Consciousness: Does Social Knowledge Affect Entry Into Consciousness?"* ; Internship Supervisors: Dr. Yair Pinto and Dr. Marte Otten.

Preprints

- Fontanesi, L., Lebreton, M., & Palminteri, S. (2018, July 20). Decomposing the effects of context valence and feedback information on speed and accuracy during reinforcement learning: A meta-analytical approach using diffusion decision modeling. Retrieved from osf.io/7k5w41

Publications

- Fontanesi, L., Gluth, S., Spektor, M. S., & Rieskamp, J. (accepted). A reinforcement learning diffusion decision model for value-based decisions. *Psychonomic Bulletin & Review.* Retrieved from osf.io/95d4p/
- Spektor, M. S., Gluth, S., Fontanesi, L., & Rieskamp, J. (in press). How similarity between choice options affects decisions from experience: The accentuation of differences model. *Psychological Review*. Retrieved from osf.io/w376r/
- Gluth, S., & Fontanesi, L. (2016). Wiring the altruistic brain. *Science*, 351(6277), 1028-1029. doi:10.1126/science.aaf4688
- van Maanen, L., Fontanesi, L., Hawkins, G. E., & Forstmann, B. U. (2016). Striatal activation reflects urgency in perceptual decision making. *NeuroImage*, 139, 294-303. doi:10.1016/j.neuroimage.2016.06.045

Talks at conferences and workshops

- **5th – 7th October 2018**: Annual Meeting of the Society for Neuroeconomics, Philadelphia, Pennsylvania, USA. Talk: "The role of dopaminergic nuclei in predicting and experiencing gains and losses: A 7T human fMRI study". Co-authors: Sebastian Gluth, Birte Forstmann, and Jörg Rieskamp
- **4th – 7th August 2016**: MathPsych (Society for Mathematical Psychology) 2016, New Brunswick, New Jersey, USA. Talk: "Modeling choices and response times during reinforcement learning". Co-authors: Sebastian Gluth, Mikhail Spektor, and Jörg Rieskamp.
- **8th – 11th June 2016**: Judgment and Decision Making Meeting, Basel, Switzerland. Talk: "Modeling response times during reinforcement learning". Co-authors: Sebastian Gluth, Mikhail Spektor, and Jörg Rieskamp.
- **4th December 2015**: Bernouilli Workshop, Basel, Switzerland. Talk: "Modeling response times during reinforcement learning". Co-authors: Sebastian Gluth, Mikhail Spektor, and Jörg Rieskamp.
- **29th – 31st July 2015**: Judgment and Decision Making Workshop, Göttingen, Germany. Talk: "Modeling response times during reinforcement learning". Co-authors: Sebastian Gluth, Mikhail Spektor, and Jörg Rieskamp.
- **8th – 11th March 2015**: TeaP (Tagung experimentell arbeitender Psychologen), Hildesheim, Germany. Talk: "Striatum codes for decision urgency". Co-authors: Leendert Van Maanen, Guy Hawkins, and Birte U. Forstmann.

Poster Presentations

- **21th – 23rd May 2018**: Symposium on Biology of Decision Making, Paris, France. Poster: "Dissociable Effects Of Valence And Information On Response Times During Reinforcement Learning". Co-authors: Maël Lebreton, and Stefano Palminteri.

- **6th – 8th October 2017**: Neuroeconomics 2017, Toronto, Canada. Poster: "Combining computational modeling and EEG to understand how reinforcement learning influences the decision process". Co-authors: Sebastian Gluth, and Jörg Rieskamp.
- **5th – 8th August 2017**: ICON conference (International Conference for Cognitive Neuroscience) 2017, Amsterdam, The Netherlands. Poster: “Modeling response times during reinforcement learning”. Co-authors: Sebastian Gluth, Mikhail Spektor, and Jörg Rieskamp.
- **21st – 25th July 2017**: MathPsych (Society for Mathematical Psychology) 2017, Warwick, UK. Poster: “Discriminating between conflict and value in foraging-like tasks: a sequential sampling-based approach”. Co-authors: Sebastian Gluth, and Amitai Shenhav.
- **28th – 30th August 2016**: Neuroeconomics 2016, Berlin, Germany. Poster: “Modeling choices and response times during reinforcement learning”. Co-authors: Sebastian Gluth, Mikhail Spektor, and Jörg Rieskamp.
- **9th – 11th May 2016**: SSM (Sequential Sampling Models) Workshop, Luzern, Switzerland. Poster: “Modeling response times during reinforcement learning”. Co-authors: Sebastian Gluth, Mikhail Spektor, and Jörg Rieskamp.
- **20th – 23rd November 2015**: SJDM (Society for Judgment and Decision Making), Chicago, USA. Poster: “Modeling response times during reinforcement learning”. Co-authors: Sebastian Gluth, Mikhail Spektor, and Jörg Rieskamp.
- **9th – 15th August 2015**: Judgment and Decision Making Summer School, Nürenberg, Germany. Poster: “Modeling response times during reinforcement learning”. Co-authors: Sebastian Gluth, Mikhail Spektor, and Jörg Rieskamp.
- **14th – 16th May 2015**: Decision & Memory Workshop, Hölstein, Basel, Switzerland. Poster: “Modeling response times during reinforcement learning”. Co-authors: Sebastian Gluth, Mikhail Spektor, and Jörg Rieskamp.

Student Scholarships and Awards

- SNSF Mobility Grant to visit Birte Forstmann’s lab at the University of Amsterdam for the duration of 9 months (2017);
- Awarded by the University of Amsterdam 35 free hours to spend at the 7Tesla scanner at the Spinoza center in Amsterdam (2017);
- Scholarship for merit at the University of Trento (2012).