University of Amsterdam

Research Master Brain and Cognitive Sciences

Project Report

A model-based approach to value-based risky decision-making

1/02/2022 - 20/07/2022

Student:
Paolo Umberto Agliati
13978926

Supervisor:
Niek Stevenson

Examiner:
Birte Forstmann

Assessor:
Julia Haaf

Group:
IMCN, Integrative Model-based
Cognitive Neuroscience Research Unit

Project:
MSc Brain and Cognitive Sciences
Research Project 1

Contents

Abstract

In an effort to draw parallels between brain functioning and the downstream resulting behavior, decision-making is frequently studied, mainly thanks to the high degree of experimental control it allows. Studying this phenomenon, we focus on the concept of risk, described as the possibility that a decision will lead to an undesirable outcome. More recently, adopting a behavioral modeling approach has provided insights into the mechanisms of decision-making. A major component of most forms of decision-making (especially the ones in which risk is involved) is the learning process, which allows for the adaptation of choices where some aspects are partially or entirely unknown to the decision-making agent. Modeling decision processes in learning tasks can offer insights into the mechanisms underlying human behavior, even more so when studies are conducted with brain imaging techniques. Here we analyze commonalities and differences in two value-based risky decision-making tasks from a behavioral standpoint, using a combined model of learning and decision-making. In the near future, the project also aims to compare similarities and differences in decision-making behavior combined with fMRI analyses of relevant cortical and subcortical networks.

1 Introduction

Risky choices are choices in which one or more courses of action could lead to an unwanted outcome, for instance, a loss, or a diminished reward. They are cases of value-based decision-making where every alternative carries subjective value, as perceived by the choice-making agent. To capture the aspects of value-based choices is of great interest to uncover what drives humans in making those decisions. We designed two risky decision-making tasks, with the aim of studying how the risk element is perceived by the subjects. In both these tasks, two types of risk can be identified. The first one is the exploration-exploitation risk, based on the ability to learn the subjective value of different stimuli (Cohen et al., 2007; Mehlhorn et al., 2015). If during the task, new, previously unseen stimuli are introduced, we want to assess how much participants are inclined to exploit (prefer the old stimuli for which a perceived value is already formed) or explore (prefer the new stimuli on which the subjects have no information about). The second type of risk is the probability-based risk (the risk associated with choosing between a low probability, high reward option, and a high probability, low reward option) and rather than learning is more based on internal preferences that bias the decision-making process. In this probability-magnitude trade-off, participants might inform their choices based on the reward magnitude, an explicit feature of the proposed alternatives, or based on the probability of receiving a reward, a characteristic that can be learned about the stimuli as the task is performed (Simen et al., 2009; Hinvest and Anderson, 2010). To study both the exploration-exploitation trade-off and the probability-magnitude trade-off in our two tasks, we constructed cognitive models that combined learning and decision-making and accounted for the effects of reward perception in the tasks. In later stages, those same tasks will be carried out in a functional magnetic resonance imaging (fMRI) scanner, to assess relationships between decision-making and brain activity (d'Acremont et al., 2009). The two tasks were designed to be analogous in all aspects, yet different on one dimension to study both commonalities and differences in the associated neural activity of a risky choice. Task one is a purely value-based task for which we expect a more pronounced neural activity in the orbital frontal cortex during risky choices (Rogers et al., 1999). The second task is designed by modifying task one, creating a value-based spatial memory task, which could involve the hippocampus more (Rangel et al., 2008). However, before making assessments on brain activity, the first step is understanding the processes that drive risky decision-making with a behavioral study.

1.1 Risk-taking and learning

Learning plays a role in the decision-making process, especially when decisions are made with limited or missing information. In such contexts, subjects construct internal preferences and representations to inform their choices, based on external, often delayed feedback inputs, able to reinforce or discourage their current behavior (Frank et al., 2004; Bogacz and Larsen, 2011). The main theoretical approaches to study learning in cognition started in the field of behavioral psychology with structural models, which constituted qualitative descriptions of cognitive or perceptual processes. Famous contributions in this sense include classical conditioning (Bitterman, 2006), instrumental conditioning (Thorndike, 1927), and goal-directed learning (Rodhom and Tolman, 1950). For all three of these main theories, the advancement of computational approaches helped in developing correspondent cognitive models, able to offer a quantitative perspective on the phenomenon. The major paradigm currently available in this field is reinforcement learning, often used to model the behavior of an agent in an uncertain environment (Dayan and Niv, 2008). In these models, agents learn the utility of certain choices based on different inputs they receive, mainly reward inputs. In the study of error-driven learning (O'Doherty et al., 2017; Sutton and Barto, 2018), reinforcement learning models are often used to describe how the outcomes of a certain action drive the subjective values people hold for these actions. However, how these subjective values influence the decision-making process is often illustrated using a descriptive model, which does not offer a process-level understanding of how decisions arise from representations and past choices, a fundamental dynamic to capture in risky decision-making.

1.2 Risk-taking and decisions: evidence accumulation models

Whereas learning is often captured using reinforcement learning models, decision-making processes are often studied using evidence accumulation models (EAMs). EAMs are used to investigate decision-making because they are able to uncover three main latent constructs in the data. First, the non-decision time, described as the time taken for sensory processing and motor preparation, in this time no cognitive aspects of the choice are processed. Second, the drift rate, described as the speed of evidence accumulation in a choice. Third, the decision threshold, which determines how much information is accumulated before the choice is made. All three of these measures reflect cognitive and non-cognitive aspects that represent information processing and response caution in subjects and can be mapped onto different brain functions (Forstmann et al., 2016). The concept of evidence accumulation was first implemented in a model by Ratcliff et al., with the diffusion decision model (DDM; Ratcliff et al., 2016). In this paradigm, an accumulator of evidence moves towards one of two thresholds with a certain drift rate. These thresholds are associated with a choice, once a threshold is reached, the model makes the corresponding choice. The DDM was able to accurately account for a variety of aspects in a wide range of choice-making tasks and led the way for the development of other cognitive models (Krajbich and Rangel, 2011; Fontanesi et al., 2019). To better adapt to more complex choices (for instance cases with more than two alternatives present) different models were developed as expansions on the DDM. One of them is the linear ballistic accumulator (LBA; Brown and Heathcote, 2008), which makes it possible to efficiently account for different context effects and multi-alternative choices. Finally, racing diffusion models constitute a further expansion of the LBA, allowing for the integration of within-trial variability and, possibly, leaving out the use of between-trial variability. This source of variability was considered an issue because it can potentially introduce an unlimited amount of flexibility in the way the model generates data, and “does not provide any further explanation to how the data were generated” (Tillman et al., 2020).

1.3 A combined perspective

Since a model for learning and a process understanding of decision-making are essential to model risky decision-making, we rely on recently developed frameworks that combine reinforcement learning models with evidence accumulation models (RL-EAM; Miletić et al., 2020; Shahar et al., 2019). Therefore, in this experiment we employed the reinforcement learning advantage racing diffusion model (RL-ARD; Miletić et al., 2021), an EAM able to take the learning process into account via a reinforcement learning framework integration. Specifically, this RL-EAM proposes that the subjective values as estimated by the reinforcement learning framework drive the speed of evidence accumulation in the EAM framework. Although we are now able to describe the decision-making process and the learning involved in the tasks, our model still lacks a comprehensive account of the expected value of a choice, that allows it to explain how, in humans, the perceived magnitude and probability of reward influence the action of choosing a stimulus over another (Rieskamp, 2008; Bhatia et al., 2021). The most prominently used framework in the field of value-based decision-making is Cumulative Prospect Theory (CPT; Tversky and Kahneman, 1992). CPT is often used to assess the values people attribute to certain choices, but also provides a “choice rule” that determines how people select which of the alternatives to pick. Since the EAM in our paradigm already represents the choice rule, we omit all elements of CPT that describe the decision-making process, leaving us with a simple model that describes how the expected value of choices drives the accumulation process. Our combined model is able to capture the relevant effects of the decision-making tasks in this experiment, namely the learning component, the decision-making processes, the expected values of choices, and the action policy. This way, we can uncover how humans face two important trade-offs in risky decision-making: the exploration-exploitation trade-off (the amount of times subjects decide to choose new, unseen stimuli over known stimuli), and the magnitude-probability trade-off (how much the explicit magnitude of the reward weighs on the subject’s choices when compared to the learned, implicit probability of reward).

1.4 Aims

In this experiment we aim to make sense of how and why different areas are activated in different types of value-based risky decision-making. It is of fundamental importance to first uncover what constitutes these choices. Thus, the first aim of the study is to capture and disentangle different types of risk in value-based choices on a behavioral level. In this study, we test two different tasks with different participants, but in a future step we will let the same participants do both of these tasks and we will develop a joint model that allows us to describe both a purely value-based task and a value-based spatial memory task. This joint model will be able to describe differences and commonalities in how risk is perceived in the two different value-based tasks. Once a working model for the choice-making process is built, a similar joint modeling approach will be used to understand and compare the behavioral aspects of decision-making with the underlying neural activity, by letting participants carry out the two tasks in an MRI scanner. In the last step, we will combine that joint model with fMRI data. This will allow the model to explain parallels between brain activation in different areas with aspects of risk in decision-making. Therefore, the study has a second, future aim: to apply this model to probe decision-making processes in cortical and subcortical networks.

2 Methods

2.1 Tasks design

In both tasks, two choice options are presented simultaneously, and participants are asked to choose one or the other stimulus. Once the stimulus is chosen, a corresponding reward is given, and two new stimuli are presented. Stimuli are characterized by a symbol and a color. In subsequent trials, the same stimulus can be presented with different colors. Stimulus color encodes for reward magnitude and is explicitly indicated to the subjects, while the stimulus itself encodes for reward probability, and is kept implicit. Each choice carries a different probability of giving reward when selected and a different magnitude of such reward. The size of the reward can range from 20 to 100, in steps of 20. A certain stimulus always carries a fixed probability of reward, which, as the task progresses, can be learned by the subject. Every stimulus differs in its associated reward probability: from 0.2 to 0.9. Participants are instructed to earn as many points as possible and to always respond before the deadline of 2 s (otherwise a “too slow” signal appears and the choice is discarded). To achieve this, the task requires subjects to learn, by trial and error, which choice options are most likely to lead to rewards and to weigh that in combination with the reward magnitudes. In the experiment, 12 abstract symbols were included. The experiment was coded in Javascript. Participants performed the tasks online. After a short practice block to let them familiarize themselves with the task, participants faced the first block, in which only two stimuli were presented, both carrying over to the second block. In every subsequent block, 2 new stimuli were introduced and 2 old stimuli carried over from previous blocks. Thus, from the second to the sixth block, four different symbols were presented. The first block comprised 30 trials, the remaining blocks each comprised 50 trials. In both tasks, only risky and consistent trials are possible. In consistent trials, the stimulus with a higher probability of reward also carries a higher magnitude. In risky trials, the stimulus with a higher probability of reward is associated with a lower magnitude or vice-versa. The two tasks are designed to be analogous, in order to be comparable and explainable in the same (modeling) framework.

Task 1

Task 1 was piloted on 11 participants. Symbol colors indicate reward magnitude, which is known to the subject. Different symbols are associated with different probabilities of receiving the reward, which is learned as the task is performed.

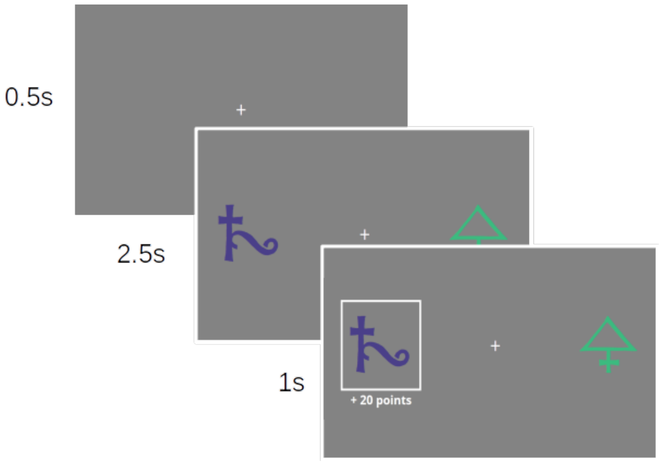

Figure 1: Task design for task 1.

Task 2

Task 2 was piloted on 26 participants. Grid colors indicate reward magnitude, which is known to the subject. The position of a highlighted cell in the grid encodes the probability of reward, which the subject has to learn. Task 2 is designed to present value-based clues which require the use of spatial memory from the subjects (King et al., 2002; Finke et al., 2008).

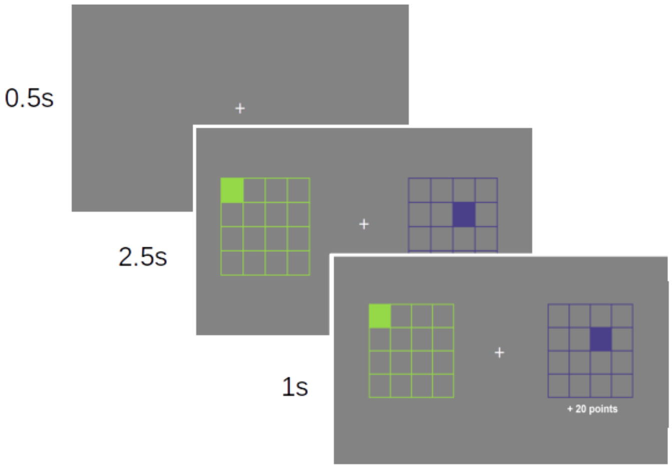

Figure 2: Task design for task 2.

2.2 Cognitive modelling

To study value-based risky decision-making in these two tasks, we developed a cognitive model which combines some aspects of reinforcement learning with the structure of EAMs, while considering the concept of the expected value from CPT.

Integrating Reinforcement Learning and EAMs

As mentioned above, value-based choices and learning are often intertwined events, and we model them together by employing the RL-ARD. The standard approach to model the learning aspect is reinforcement learning, which has been integrated into other, more complex models of choice (Fontanesi et al., 2019; Miletić et al., 2021). In classical reinforcement learning, learning agents update their perceived value of the choice according to the reward they get, the learning rate, and an action selection policy (Sutton and Barto, 2018). The most prominent reinforcement learning paradigm is Q-learning (Watkins and Dayan, 1992), in which Q-values represent the perceived utility of the choice and are updated using a delta update rule:

$$Q_{i,t+1} = Q_{i,t} + \alpha(r_t - Q_{i,t}) \tag{1}$$

where Q_i, t is the value representation of choice option i on trial t, r_t is the reward on trial t, and α is the learning rate, that modulates the changes in Q-values in response to a prediction error (defined as the difference between actual and perceived reward: $r_t - Q_i, t$). For our tasks, Q-learning is very useful because it can approximate the probability of reward as perceived by the learning subjects. This also means that Q-values can be used to assess whether our model is learning the values of the proposed alternatives, and if so, whether this learning process resembles the one happening in human participants. To better account for all the interacting facets of a choice, RL-ARD also introduces the advantage framework which, when compared to the simpler racing diffusion model, allows to attribute different aspects of the choice to different accumulators that compete towards a common threshold (van Ravenzwaaij et al., 2019; Miletić et al., 2021). Drift rates for the RL-ARD are calculated as follows:

$$\begin{aligned} dx_1 &= [V_0 + w_d(Q_1 - Q_2) + w_s(Q_1 + Q_2)]dt + sW \\ dx_2 &= [V_0 + w_d(Q_2 - Q_1) + w_s(Q_1 + Q_2)]dt + sW \end{aligned} \tag{2}$$

Where dx are the drift rates, V_0 represents the evidence-independent speed of accumulation, w_d is the difference (advantage) of the evidence for one choice option over the other, w_s is the sum of the total available evidence and sW is the standard deviation of within-trial noise. Finally, Q_n in Equation 2 represent Q-values for different choices, and are updated via the simple delta update rule (see Equation 1).

We then integrated the expected value of each choice from CPT, described as magnitude times probability of reward and we adapted them to our RL-ARD model the following way:

$$\begin{aligned} dx_1 &= [V_0 + w_d(\gamma_1 - \gamma_2) + w_s(\gamma_1 + \gamma_2)]dt + sW \\ dx_2 &= [V_0 + w_d(\gamma_2 - \gamma_1) + w_s(\gamma_1 + \gamma_2)]dt + sW \end{aligned} \tag{3}$$

where:

$$\begin{aligned} \gamma_1 &= M * Q_1 \\ \gamma_2 &= M * Q_2 \end{aligned} \tag{4}$$

with M being the reward magnitude itself, and Q_1 and Q_2 being the Q-values (expected probability of reward), updated via the delta update rule (see Equation 1).

Model Parameters Estimation

In the framework used for this experiment, all the parameters that constitute the reinforcement learning aspect are estimated simultaneously with all the other evidence accumulation parameters, via hierarchical Bayesian inference (Lee, 2011). Estimation of the parameters is carried out via Markov-Chain Monte Carlo (MCMC) processing, which generates random samples drawn from a distribution. In this sampling method, new samples are usually drawn from a normal distribution and do not depend on any samples before the previous one (van Ravenzwaaij et al., 2018). Among the advantages of this procedure, it allows one to avoid characterizing the distribution analytically and to make the parameters converge to the posterior. At each iteration, a new proposal is generated from a proposal distribution, and the likelihood and posterior are calculated via Bayes rule:

$$P(\theta|X) = \frac{P(X|\theta)\, P(\theta)}{P(X)} \tag{5}$$

Where the values of P(θ | X) are the probability of the parameter values given the experimental data, which constitute the posterior distribution. The distribution is built using Bayes rule by multiplying a set of parameters priors P(θ) times the likelihood of the data given such parameters P(X | θ). To form the posterior, such value is then weighted on the probability of the data itself occurring, P(X). For our purposes, P(X) can be considered a normalizing constant that allows the posterior values to be included in the [0,1] interval (and therefore to obtain a probability distribution as an output) and does not need to be estimated in the MCMC framework. Usually, an acceptance policy on the new parameters proposals is implemented to avoid local minima, so that over time the posterior for the model parameters is reached. ”At each iteration, if the new proposal is accepted, it becomes the next sample in the MCMC chain, otherwise the next sample in the MCMC chain is just a copy of the most recent sample” (van Ravenzwaaij et al., 2018). The usage of a Bayesian approach also gives a substantial advantage in treating the data when compared, for example, to null hypothesis significance testing (NHST) with p-value in classical statistics, since a Bayesian perspective offers internal consistency, independence from model complexity, and significant advantages when it comes to hypothesis testing; for an in-depth discussion, see Wagenmakers et al., 2018. In our study, these techniques are applied in the construct of a hierarchical model. Hierarchical sampling allows generating a distribution among subjects’ sampled parameters, forming the so-called ”group level”, and from such distribution draw samples to inform the priors in the next iteration at the individual level. This greatly aids the efficiency with which the model converges towards the posterior parameter values, and helps describe common effects underlying the modeled task (Vandekerckhove et al., 2011). In our model, the parameters are estimated via particle metropolis within Gibbs (PMwG) sampling. This sampling technique implements Gibbs sampling (van Ravenzwaaij et al., 2018) to shape the group level and draw better parameter values from it. PMwG constructs the group level using a multivariate normal distribution that is built from each parameter’s variance and all the covariances between the parameters. This allows accounting for combined effects stemming from the interplay of different elements of the choice. Furthermore, PMwG uses three stages of sampling: the burn-in phase, in which parameters are still far from the posterior and will be discarded; the adaptation phase, in which parameter values will be used as a starting point to converge on the posterior; and finally the sampling phase, in which parameters converge towards the posterior. Since joint modeling of brain and behavioral data will be needed for this experiment, accounting for relations between parameters might be extremely useful. Thus, PMwG constitutes a valuable tool in this regard, even for subsequent experiments.

3 Results

3.1 Capturing the magnitude-probability trade-off

In this study, we carried out two risky decision-making experiments. Task 1 is a purely value-based task, tested on 11 subjects, while Task 2 is a value-based spatial memory task on which we tested 26 subjects. We used the same model architecture to describe both of the tasks. We first assessed how well the model accounts for the behavioral data by comparing the accuracy and response speed (RT) of human participants and the model in two conditions of risky and consistent trials (Figure 3). For accuracy, we analyzed the mean, while for response speed the median, the 10th percentile, and the 90th percentile were analyzed. While the median summarizes the central tendency, the differences between the 10th, 50th, and 90th percentiles summarize variability in the response times data for the factor of interest. The data is represented for both task 1 (Symbol task, Figure 3A) and task 2 (Grid task, Figure 3B). In consistent trials, the stimulus with a higher reward also had a higher probability of giving reward when selected, whereas in risky trials this was not the case, and the magnitude of the stimuli pointed to a different response than the associated probability of the stimuli. Therefore, it was also harder for the participants to find the most rewarding stimulus. Harder choices are usually associated with lower accuracy and higher response speed compared to correct choices (Smith and Ratcliff, 2004). As shown, the model provides a good description of this effect. As mentioned in section 2.1, the task design allows only risky and consistent trials, which in Figure 3 are referred to as "Trial Type". For the accuracy over trial type measure, the model seems to underestimate the accuracy of consistent trials and overestimate the accuracy of risky trials for both tasks. Human participants tended to show a more clear difference in accuracy between risky and consistent trials.

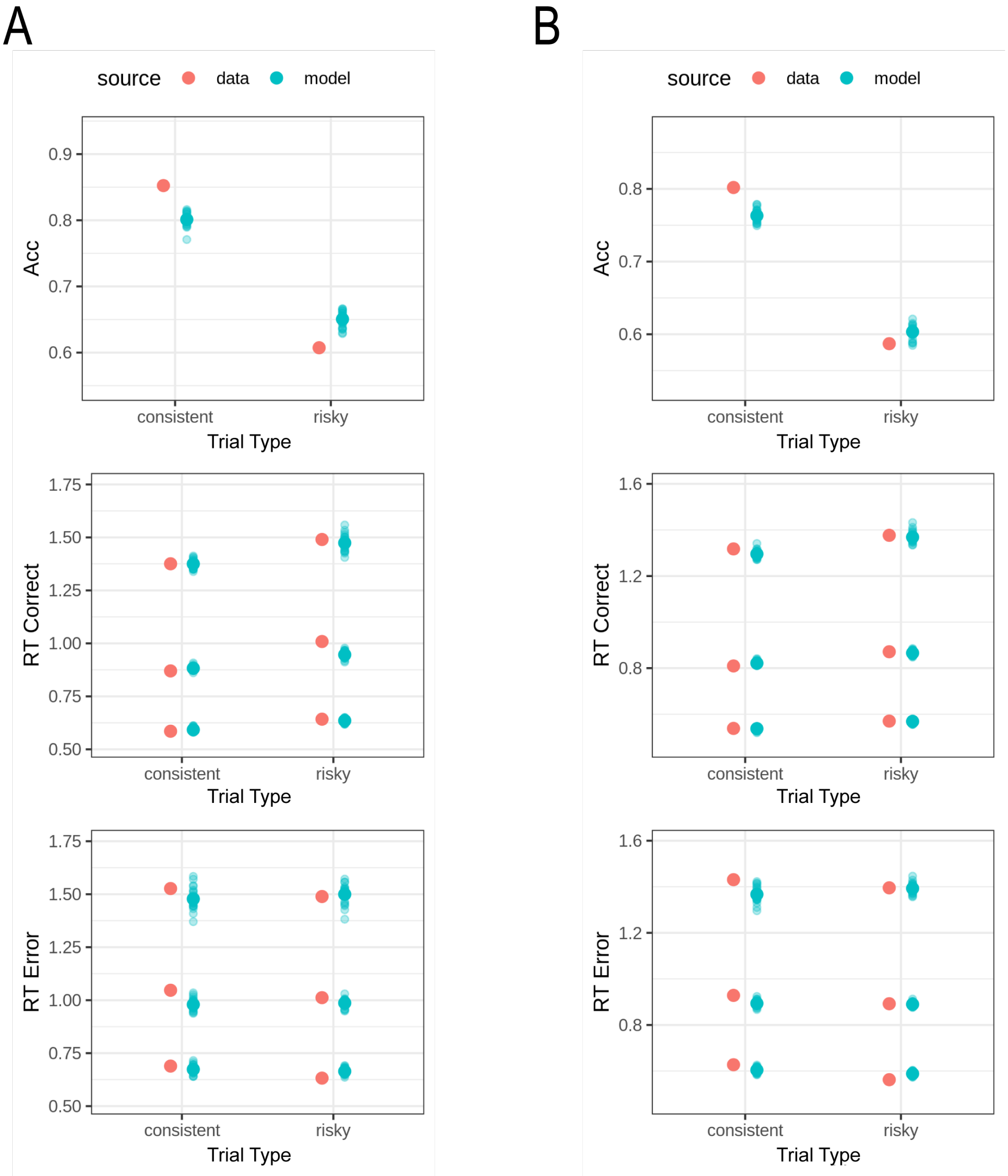

Figure 3: Trial type over accuracy and response speed (for correct and incorrect responses) in task 1 (A) and task 2 (B). For RTs, the 10th, 90th percentile, and median values are shown.

We further looked at stimuli of different magnitude and the percentage of times such stimuli were chosen. Stimulus magnitude is a measure of how participants approach the magnitude-probability trade-off and can be used to assess if the model accounts for participants' tendencies. In Figure 4, magnitude of reward is compared with the percentage of chosen stimulus and response speed. Data are plotted for both task 1 (Figure 4A) and task 2 (Figure 4B). RTs are well accounted for by the model in the 10th percentile, median value, and 90th percentile.

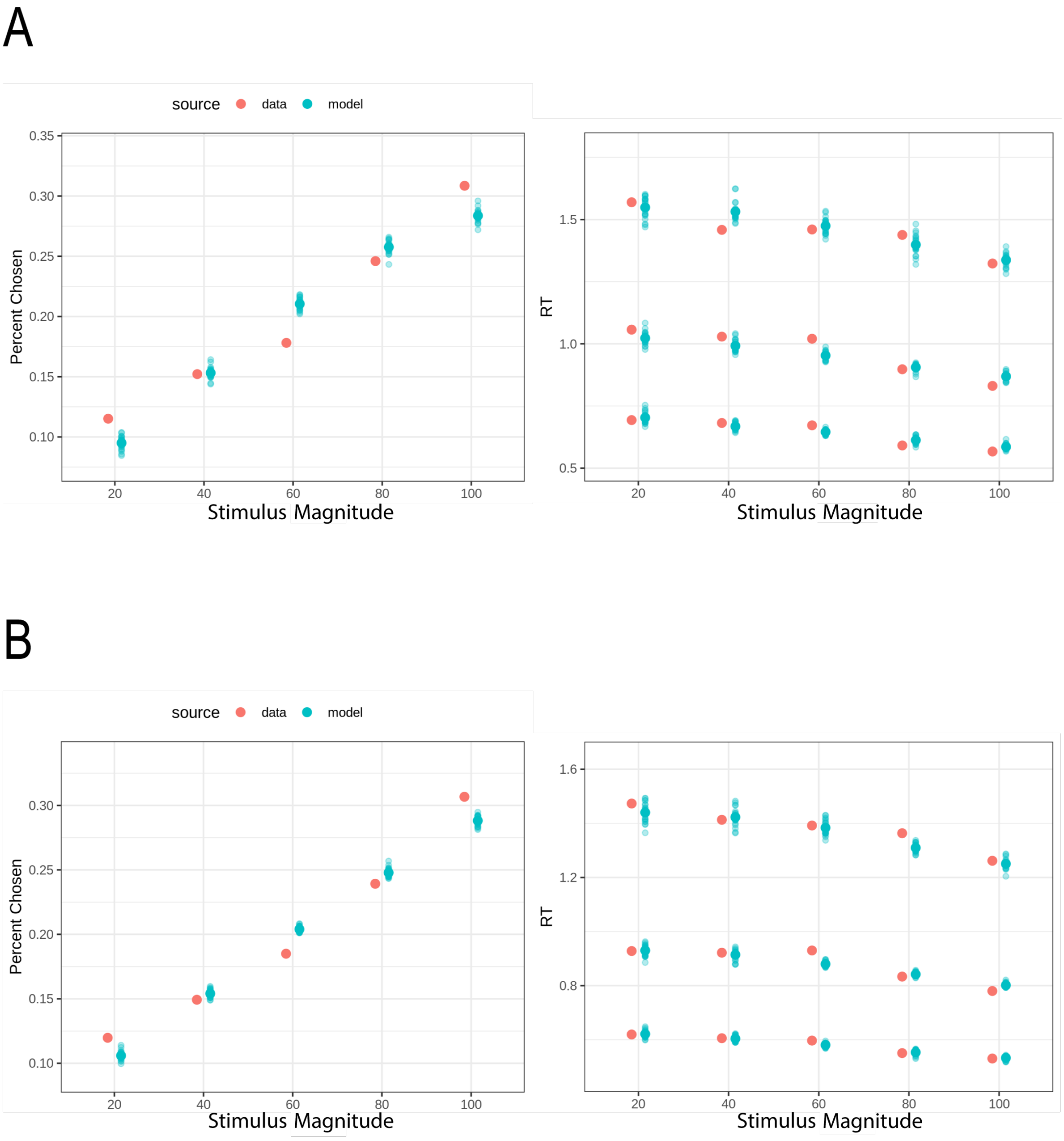

Figure 4: Stimulus Magnitude over the percentage of times that stimulus is chosen and RTs for task 1 (A) and task 2 (B).

3.2 Capturing the learning component

In the experiment, we were also interested in to what extent the model could account for the learning elements of the tasks. For both tasks, participants tended to gradually learn the probabilities associated with the stimuli and to choose higher probability stimuli as the task progresses. As can be seen in Figure 5, the model accounts for the learning effect both in response speed and percentage of stimuli chosen. As shown, participants tended to choose less stimuli carrying a lower probability of reward, showing that they learned the probabilities associated with the stimuli. In both tasks, the model follows this tendency. For this measure, the model slightly underestimates the response accuracy of high probability stimuli and overestimates the accuracy of low probability stimuli. RTs tend to be slightly higher for lower probability stimuli since choosing such stimulus constitutes a harder choice. The model follows this trend for the median, 10th percentile, and 90th percentile values.

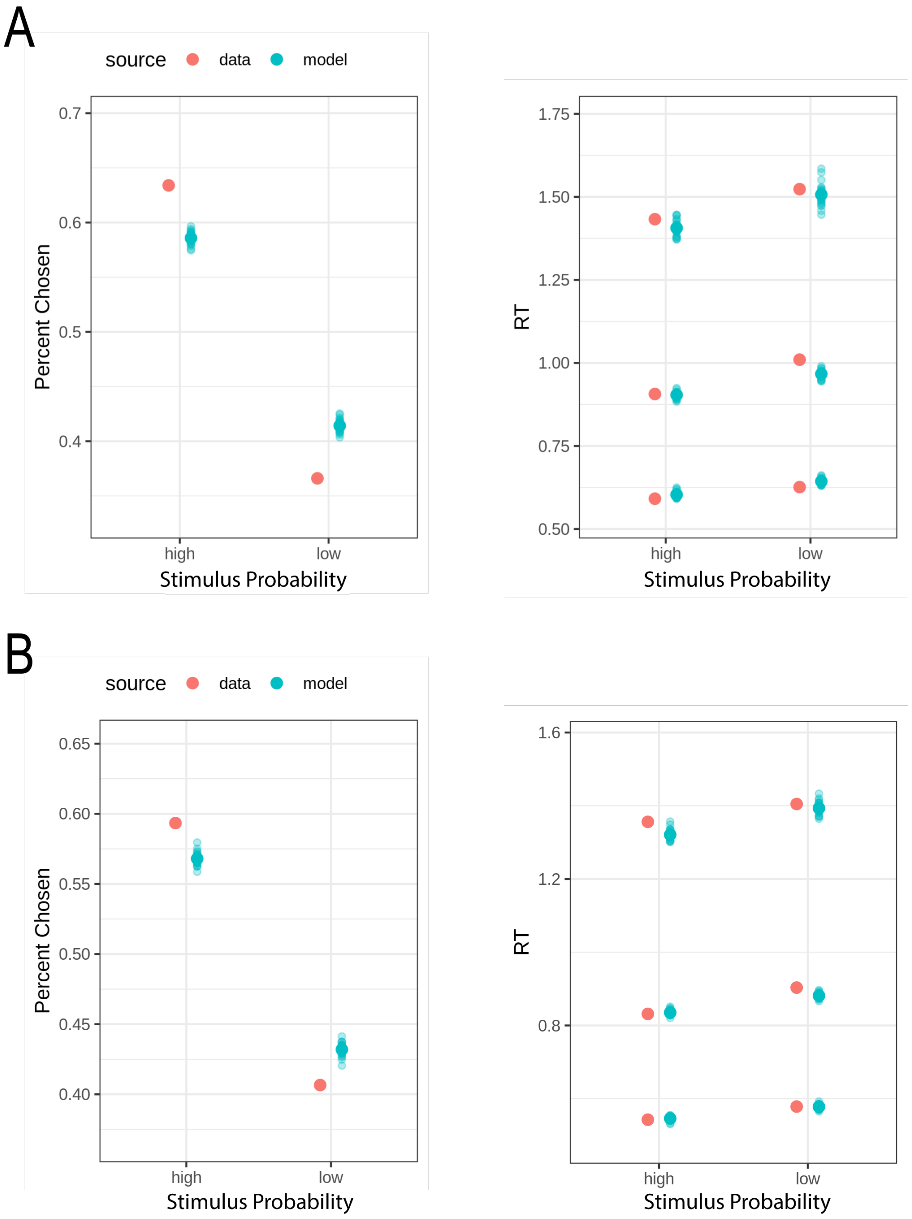

Figure 5: Stimulus probability (probability of receiving reward) is divided into low probability (below 0.5) and high probability (above and including 0.5) and plotted over the percentage of times that stimulus is chosen and the RTs in task 1 (A) and task 2 (B).

Next, we tested the effect of stimulus novelty on accuracy and response speed. During the experiment, participants were sometimes faced with two stimuli they already encountered, with two stimuli newly introduced in the current block, or with a mixed case, in which just one of the two alternatives is new, and the other one is known to the subject. The model accounts well for RTs in all three of these cases in the median, 10th percentile, and 90th percentile values. Interestingly, a different behavior between the model and the participants can be seen for accuracy in both tasks (Figure 6A for task 1 and Figure 6B for task 2), especially when both stimuli are new. In task 1, participants seem to have higher accuracy for old stimuli, an intermediate accuracy for new stimuli, and a lower accuracy for mixed stimuli. For task 1, response speeds follow an inverse trend: they are higher in mixed stimuli and decrease in new and old stimuli. This tendency, however, is lost in task 2, where no significant difference can be found in either accuracy or response speed relative to stimulus novelty.

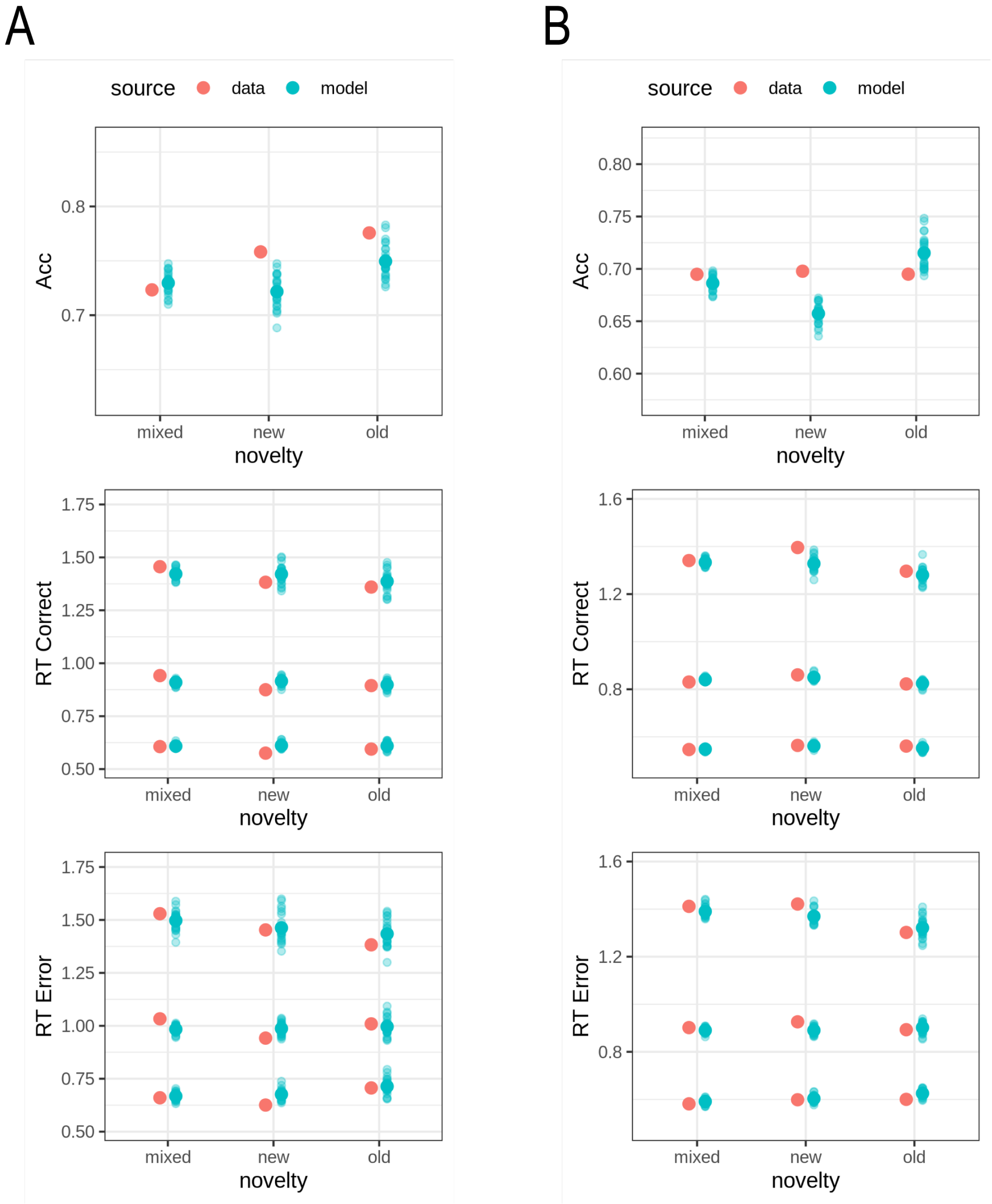

Figure 6: Stimulus novelty (old, mixed, and new choices) over accuracy of choice and RTs for task 1 (A) and task 2 (B).

3.3 Parameter estimates

We extracted the group-level model's parameter estimates for both tasks. Posterior parameter values are comparable in the two tasks, although the posterior parameter distribution for task 1 is slightly wider than the one of task 2. Such effect could be a consequence of the lower number of participants task 1 was tested on (11 compared to the 26 of task 2). In table 1 the 2.5, 97.5 percentile, and the median parameter values for every estimated parameter in task 1 are reported.

		Percentiles		
		2.5%	**50%**	**97.5%**
Parameters	**t0**	0.02	0.10	0.18
	V0	1.27	1.60	1.86
	B	1.86	2.32	2.65
	aV	0.02	0.06	0.14
	wD	1.25	2.21	2.96
	wS	0.20	0.38	0.60
	Q0	0.31	0.46	0.63

Table 1: Parameter estimates for task 1. t0: non-decision time; V0: evidence-independent speed of accumulation; B: decision threshold; aV: learning rate; wD: advantage of the evidence for one choice option over the other; wS: sum of the total available evidence; Q0: Q-values initial estimates.

In table 2 the 2.5, 97.5 percentile, and the median parameter values for every estimated parameter in task 2 are reported.

		Percentiles		
		2.5%	**50%**	**97.5%**
Parameters	**t0**	0.06	0.09	0.13
	V0	1.39	1.52	1.66
	B	1.86	2.15	2.43
	aV	0.02	0.04	0.08
	wD	1.72	2.09	2.47
	wS	0.42	0.49	0.57
	Q0	0.27	0.37	0.47

Table 2: Parameter estimates for task 2. t0: non-decision time; V0: evidence-independent speed of accumulation; B: decision threshold; aV: learning rate; wD: advantage of the evidence for one choice option over the other; wS: sum of the total available evidence; Q0: Q-values initial estimates.

4 Discussion

4.1 Conclusion

In this experiment, we studied the influence of risk on decision-making in two different tasks: a value-based task and a spatial memory value-based task. To that end, we developed a model that could account for the effects of both tasks, highlighting the differences and commonalities between them. On the one hand, the model captures the differences between risky and consistent choices, and accounts for the probability, magnitude and learning effects in our design well.
On the other hand, the model still needs improvement in a few aspects. Across most of the measures, the model fits the data of task 2 (spatial memory task) better than task 1 (symbol task). This might be due to the fact that task 2 was tested on more participants. Furthermore, only some aspects of the task can be directly mapped onto the model parameters. This issue is of particular relevance as we want the model to explain decision-making as clearly as possible,

and therefore we want its parameters to be able to directly tell something about the studied phenomenon. An example of such direct mapping is the parameter Q0, which is the initial estimate of Q-values and it maps onto the tendency to favor exploration over exploitation. A higher Q0 value will favor choosing new, unseen stimuli more compared to a lower Q0 estimate, meaning that Q0 can represent a measure of how people value new choices. To better appreciate why that is important, let us consider the measures of novelty in Figure 6. From the data, there seems to be a bias in novelty for task 1 over task 2: in task 1 introducing new symbols affects the accuracy more than introducing new grid layouts in task 2. Even though the model as it is now, does not account for this novelty bias (the model predicts a similar change in accuracy for both tasks), a direct measure of the value given to new stimuli can be found in the model parameter Q0. This means in future steps we will be able to directly assess and quantify if this novelty bias will be better captured. However, some other aspects of the task, like the magnitude-probability trade-off, do not map directly onto the model parameters in a similar way. This means that in the future, when developing a joint model, a parameter that accounts for this effect will be needed, to allow to directly quantify to which extent the trade-off is captured.

4.2 Future outlook

With this experiment, we managed to set the basis to investigate risk in joint models of value-based decision-making. The modeling approach we adopted allows us to explain the elements of risk in a choice, and can be used to expand on how risk is encoded in different brain areas during the decision-making process. However, more work is needed to meet the initial aims of the study. The first and closest future step will be to estimate a joint model for the behavioral data of both tasks since for now different participants were called for the two tasks in the piloting. Having the same participants complete both of the tasks and run the joint model on such data would give a clearer idea of the feasibility of the fMRI experiment, and would possibly highlight new challenges to overcome before testing the participants in the scanner. Once we develop a joint model that accounts for the effects in both of the behavioral tasks, we run the experiment and start collecting data for both tasks in an MRI scanner, so that we can obtain brain and behavioral data of all participants. Finally, starting from the existing EAM, we can develop a joint model that accounts for brain and behavioral data, and we can draw parallels between activation in certain areas and different kinds of tasks. We might expect a similar activation of brain networks classically linked to decision-making to be shared between both tasks; for instance, the connections between the basal ganglia– dopamine (BG-DA) system and the orbitofrontal cortex (OFC) or the projections from these areas to the amygdala (Frank and Claus, 2006). We might also expect a more significant activation of the OFC for the purely value-based (symbol) task, and a stronger hippocampal response for the spatial memory (grid) task. Thanks to this joint model, we would also be able to draw relations between brain activation and different behavior among subjects in the same task, thus uncovering how cortical and subcortical structures are involved in value-based risky decision-making.

References

Bhatia, S., Loomes, G., & Read, D. (2021). Establishing the laws of preferential choice behavior. *Judgment and Decision Making*, *16*(6).

Bitterman, M. E. (2006). Classical conditioning since Pavlov. https://doi.org/10.1037/1089-2680.10.4.365

Bogacz, R., & Larsen, T. (2011). Integration of Reinforcement Learning and Optimal Decision-Making Theories of the Basal Ganglia. *Neural Computation*, *23*(4), 817–851. https://doi.org/10.1162/NECO_a_00103

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3). https://doi.org/10.1016/j.cogpsych.2007.12.002

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481). https://doi.org/10.1098/rstb.2007.2098

d'Acremont, M., Lu, Z.-L., Li, X., Van der Linden, M., & Bechara, A. (2009). Neural correlates of risk prediction error during reinforcement learning in humans. *NeuroImage*, *47*(4), 1929–1939. https://doi.org/10.1016/j.neuroimage.2009.04.096

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The Good, The Bad and The Ugly. https://doi.org/10.1016/j.conb.2008.08.003

Finke, C., Braun, M., Ostendorf, F., Lehmann, T. N., Hoffmann, K. T., Kopp, U., & Ploner, C. J. (2008). The human hippocampal formation mediates short-term memory of colour-location associations. *Neuropsychologia*, *46*(2). https://doi.org/10.1016/j.neuropsychologia.2007.10.004

Fontanesi, L., Gluth, S., Spektor, M. S., & Rieskamp, J. (2019). A reinforcement learning diffusion decision model for value-based decisions. *Psychonomic Bulletin and Review*, *26*(4). https://doi.org/10.3758/s13423-018-1554-2

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E. J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual Review of Psychology*, *67*. https://doi.org/10.1146/annurev-psych-122414-033645

Frank, M. J., & Claus, E. D. (2006). Anatomy of a decision: Striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological Review*, *113*(2). https://doi.org/10.1037/0033-295X.113.2.300

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

Hinvest, N. S., & Anderson, I. M. (2010). The effects of real versus hypothetical reward on delay and probability discounting. *Quarterly Journal of Experimental Psychology*, *63*(6). https://doi.org/10.1080/17470210903276350

King, J. A., Burgess, N., Hartley, T., Vargha-Khadem, F., & O'Keefe, J. (2002). Human hippocampus and viewpoint dependence in spatial memory. https://doi.org/10.1002/hipo.10070

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(33). https://doi.org/10.1073/pnas.1101328108

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*(1). https://doi.org/10.1016/j.jmp.2010.08.013

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., Hausmann, D., Fiedler, K., & Gonzalez, C. (2015). Unpacking the exploration-exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, *2*(3). https://doi.org/10.1037/dec0000033

Miletić, S., Boag, R. J., & Forstmann, B. U. (2020). Mutual benefits: Combining reinforcement learning with sequential sampling models. *Neuropsychologia*, *136*. https://doi.org/10.1016/j.neuropsychologia.2019.107261

Miletić, S., Boag, R. J., Trutti, A. C., Stevenson, N., Forstmann, B. U., & Heathcote, A. (2021). A new model of decision processing in instrumental learning tasks. *eLife*, *10*. https://doi.org/10.7554/eLife.63055

O'Doherty, J. P., Cockburn, J., & Pauli, W. M. (2017). Learning, Reward, and Decision Making. https://doi.org/10.1146/annurev-psych-010416-044216

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, *9*(7), 545–556. https://doi.org/10.1038/nrn2357

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. https://doi.org/10.1016/j.tics.2016.01.007

Rieskamp, J. (2008). The Probabilistic Nature of Preferential Choice. *Journal of Experimental Psychology: Learning Memory and Cognition*, *34*(6). https://doi.org/10.1037/a0013646

Rodhom, C., & Tolman, E. C. (1950). Purposive Behavior in Animals and Men. *The American Journal of Psychology*, *63*(2). https://doi.org/10.2307/1418946

Rogers, R. D., Owen, A. M., Middleton, H. C., Williams, E. J., Pickard, J. D., Sahakian, B. J., & Robbins, T. W. (1999). Choosing between small, likely rewards and large, unlikely rewards activates inferior and orbital prefrontal cortex. *Journal of Neuroscience*, *19*(20). https://doi.org/10.1523/jneurosci.19-20-09029.1999

Shahar, N., Hauser, T. U., Moutoussis, M., Moran, R., Keramati, M., Consortium, N. S., & Dolan, R. J. (2019). Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. *PLoS Computational Biology*, *15*(2). https://doi.org/10.1371/journal.pcbi.1006803

Simen, P., Contreras, D., Buck, C., Hu, P., Holmes, P., & Cohen, J. D. (2009). Reward Rate Optimization in Two-Alternative Decision Making: Empirical Tests of Theoretical Predictions. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6). https://doi.org/10.1037/a0016926

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. https://doi.org/10.1016/j.tins.2004.01.006

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An Introduction (2nd edition 2018)* (Vol. 3).

Thorndike, E. L. (1927). The Law of Effect. *The American Journal of Psychology*, *39*(1/4). https://doi.org/10.2307/1415413

Tillman, G., Van Zandt, T., & Logan, G. D. (2020). Sequential sampling models without random between-trial variability: the racing diffusion model of speeded decision making. https://doi.org/10.3758/s13423-020-01719-6

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4). https://doi.org/10.1007/BF00122574

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical Diffusion Models for Two-Choice Response Times. *Psychological Methods*, *16*(1). https://doi.org/10.1037/a0021765

van Ravenzwaaij, D., Brown, S. D., Marley, A. A., & Heathcote, A. (2019). Accumulating Advantages: A New Conceptualization of Rapid Multiple Choice. *Psychological Review*. https://doi.org/10.1037/rev0000166

van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov Chain Monte–Carlo sampling. *Psychonomic Bulletin and Review*, *25*(1). https://doi.org/10.3758/s13423-016-1015-8

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin and Review*, *25*(1). https://doi.org/10.3758/s13423-017-1343-3

Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3-4). https://doi.org/10.1007/bf00992698