FACULTY OF PSYCHOLOGY AND EDUCATIONAL SCIENCES

COMPUTATIONAL MODEL COMPARISON IN THE TWO-STEP DECISION-MAKING TASK

Word count: 12,159

Frederik De Spiegeleer
Student number: 01204797

Supervisor(s): Dr. Elise Lesage, Prof. Dr. Tom Verguts

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Master of Science in Psychology, main subject Theoretical and Experimental Psychology

Academic year: 2020 - 2021

GHENT UNIVERSITY

Acknowledgement

First of all I would like to thank Dr. Elise Lesage for her guidance and the many meetings we had to discuss this thesis. You introduced me to a field of research that will remain an interest of mine. In addition, you gave me the opportunity to carry out computational modeling in practice and I would like to continue modeling because of this experience. Thank you for letting me explore and express myself during these last two years.

Secondly, I would like to express my gratitude to Prof. Dr. Tom Verguts. You have not only introduced us experimental psychology students to programming and computational modeling during your courses over the past few years, but you have also made these complex topics tangible. In addition, during our meetings you also provided clarification where necessary, which really guided me through this complex matter. Thank you for given me valuable skills for my future career.

Thirdly, I would like to thank Prof. Dr. Qi Chen and Prof. Dr. Tingyong Feng for collecting this large sample size and allowing me to use this data for my master's thesis.

And finally, I also want to express my gratitude to my loved ones. Thank you, my friends, for reminding me that life has much more meaning than we sometimes think there is and for having interesting discussions with me without any limitations. Thank you, my love, for bringing color and adventure into my life, for letting me be myself, and for being the mental backbone when I needed you. And, of course, thank you, family, I could never have started or finished this without all the help you have given me.

Abstract

Goal-directed and habitual behavior are two fundamental forms of behavior in organisms, but their relative importance remains unclear. In reinforcement learning, goal-directed and habitual behavior are referred to as model-based and model-free responses, respectively. An extensively used decision-making task, called the two-step task, allows capturing a balance between model-based and model-free systems, assuming humans use a mixture of these systems. However, a recent study demonstrated that clear instructions cause participants to mainly use model-based strategies. Moreover, previous research suggested that certain behaviors could be better understood at a different dimension of learning and decision-making. In the current study, participants performed a modified two-step task in which model-based strategies are more accurate than model-free ones, unlike the original task. Here, we examined whether participants also mainly use model-based responses with improved instructions, and how the trade-off between exploration and exploitation drives our behavior in the decision-making process. The results of a model comparison showed that the subjects mainly used model-based control, explored more when the stimuli changed, and had high learning rates in our task. Furthermore, as our task was modified to maximize the benefits of model-based control, we did not find a relationship between model-based control and accuracy, but instead observed that more exploitation was related to higher accuracy. Our results suggest that the distinction between model-based and model-free learning is not sufficient to understand behavior in the two-step task. We address that future research should consider several factors when using the two-step task.

Nederlandstalige Samenvatting

Doelgericht en gewoontegedrag zijn fundamentele vormen van gedrag, maar hun relatieve belang blijft onduidelijk. Bij versterkend leren (*reinforcement learning*) worden doelgericht en gewoontegedrag respectievelijk modelgebaseerde en modelvrije reacties genoemd. Een veelgebruikte besluitvormingstaak, de tweestapstaak genoemd, maakt het mogelijk om een balans te vinden tussen modelgebaseerde en modelvrije systemen, ervan uitgaande dat mensen een combinatie gebruiken. Een recent onderzoek toonde echter aan dat duidelijke instructies ervoor zorgen dat deelnemers voornamelijk modelgebaseerde strategieën gebruiken. Bovendien suggereerde voormalig onderzoek dat bepaalde gedragingen beter begrepen worden op een andere dimensie van leren en besluitvorming. In de huidige studie voerden de deelnemers een aangepaste tweestapstaak uit waarin modelgebaseerde strategieën accurater zijn, in tegenstelling tot de oorspronkelijke taak. Hier hebben we onderzocht of de deelnemers ook voornamelijk modelgebaseerde reacties gebruiken met verbeterde instructies, en hoe de afweging tussen exploratie en exploitatie onze beslissingen aanstuurt. De resultaten toonden aan dat de proefpersonen voornamelijk modelgebaseerde controle gebruikten, meer exploreerden wanneer stimuli veranderden en een hoge leersnelheid hadden in onze taak. Bovendien, aangezien de taak werd aangepast om modelgebaseerde controle voordelig te maken, vonden we geen verband tussen modelgebaseerde controle en prestatie, maar zagen we dat meer exploitatie gerelateerd was aan betere prestatie. Onze resultaten laten zien dat het onderscheid tussen modelgebaseerd en modelvrij leren niet voldoende is om gedrag in de tweestapstaak te begrijpen. We bespreken dat toekomstig onderzoek verschillende factoren in overweging moet nemen bij het gebruik van de tweestapstaak.

Table of Contents

Computational Model Comparison in the Two-Step Decision-Making Task

It is generally accepted that organisms can learn to choose actions in an environment by experiencing outcomes of their actions (reinforcement and punishments; i.e., instrumental learning; Skinner, 1938). The basic idea is that rewarded actions are more likely to be taken in the future while punished actions are less likely to be repeated (Thorndike, 1898). So, organisms decide to take actions based on the learned associations between the outcomes of these actions and the actions themselves. However, to learn this in an optimally efficient manner, organisms must learn about much more than just associations between stimuli, responses, and outcomes. For example, organisms can also learn about their environment without experiencing immediate reward or punishment, and this information can be used to make future decisions as well (Gershman & Niv, 2010; Tolman, 1948).

While many organisms are able to carefully consider the possible outcomes of their actions by using information they learned about their environment, they often simply repeat the actions that previously resulted in a desirable outcome. This is often referred to as a distinction between goal-directed and habitual behavior. Reinforcement learning is a computational framework used to distinguish goal-directed behavior from habitual behavior. Goal-directed behavior is cognitively flexible behavior that leverages an internal model of the environment. Therefore, in reinforcement learning, goal-directed behavior is referred to as using *model-based* strategies; using strategies based on a “model of the world” (Dayan & Niv, 2008). A model-based strategy uses information learned from the environment in addition to the outcomes associated with actions. Thus, model-based strategies are more accurate but need a higher mental effort as well (Daw, Niv, & Dayan, 2005). On the other hand, habitual behavior is referred to as using *model-free* strategies in reinforcement learning (Dayan & Niv, 2008), because it simply uses outcome associations of actions without taking the model of the environment into account. Therefore, model-free strategies are less accurate but also less effortful to use (Daw et al., 2005).

For instance, imagine you want to buy a salad. One possibility is that you choose to buy the salad from the supermarket you always go to to get your groceries (i.e., a model-free strategy). However, another possibility is to choose the shop two blocks away, in which you know the prices of the salad are lower (i.e., a model-based strategy). So, by deliberating how the environment works (e.g., prices for a salad at different places), organisms are able to make better decisions. Instead, organisms often tend to repeat their previous decisions although it was not the best option (e.g., going to the same supermarket).

Daw, Gershman, Seymour, Dayan, and Dolan (2011) introduced the two-step task, a behavioral paradigm that operationalises the distinction between the model-based and model-free systems. This paradigm has since been extensively used, especially to make a distinction between the neural substrates of the two systems (e.g., Doll, Duncan, Simon, Shohamy, & Daw, 2015; Lee, Shimojo, & O'Doherty, 2014; Piray, Toni, & Cools, 2016), and to examine the effects of working memory capacity and stress on the balance between the two systems (e.g., Otto, Gershman, Markman, & Daw, 2013; Otto, Raio, Chiang, Phelps, & Daw, 2013; Radenbach, Reiter, Engert, Sjoerds, Villringer, Heinze, Deserno, & Schlagenhauf, 2015). The two-step task is a decision-making task designed to capture a parameter (by the use of computational modeling) that represents the balance between model-free and model-based learning. In other words, it allows capturing a parameter that represents the balance between behavior using either simply outcome associations or behavior that also takes a model of the task into account.

Daw et al. (2011) demonstrated that humans are not just pure model-free or pure model-based learners, but seem to use a mixture of these two systems during the two-step task. Using a mixture of the two systems actually means that an agent learns the outcome associated with an action partially by using a model of the task. Thus, the more an agent uses the model-free system, the more the agent relies on actual observed outcomes to learn an action's expected outcomes. The more an agent uses the model-based system, the more the agent also updates the expected outcomes of an action based on the structure of the task.

However, certain behaviors in the two-step task can not be understood by simply making a distinction between model-based and model-free learning (Collins & Cockburn, 2020; Daw, 2018). Therefore, the current study investigated if there are other ways we can understand the behavior in the two-step task. Before we explain how we examined this, we will first elaborate on how the two-step task works and explain some of its limitations.

The Two-Step Task

The two-step task by Daw et al. (2011) is a decision-making task in which each trial consists of two step phases. In the first-step phase of a trial, participants have to choose between two stimuli (Figure 1A). Each stimulus in the first-step phase has a probability of leading to one of two different second-step phases. In each second-step phase, participants have to choose again between two stimuli. In these second-step phases, each choice is associated with a different probability of receiving a binary reward (i.e., receiving a reward or not). The reward probabilities of the stimuli in the second-step phases fluctuate slowly and independently so that the participants have to learn throughout the task (Figure 1B).

Importantly, each choice in the first-step phase has a common transition (i.e., a 70% chance that it leads to one of the second-step phases) and a rare transition (i.e., a 30% chance that it leads to the other second-step phase); the common transition of one choice is the rare transition of the other choice. These probability transitions of the first-step phase are crucial for dissociating between model-free and model-based learning. For example, a model-free learner will more likely choose a stimulus in the first-step phase that led to a reward in the previous trial, regardless of the probability transition (common or rare transition) because the model-free learner does not take the (transition) structure of the task into account. In contrast, a model-based learner will less likely repeat a decision in the first-step phase after a reward with a rare transition, because the likelihood of leading to the previous second-step phase is higher when choosing the other option in the first-step phase. In addition, the model-based strategy will more likely repeat a decision after an unrewarded rare transition since that choice will more likely be followed by the other second-step phase. With the

latter in mind, the two-step task can estimate a weighting parameter that establishes the balance between model-free and model-based learning of the behavioral data.

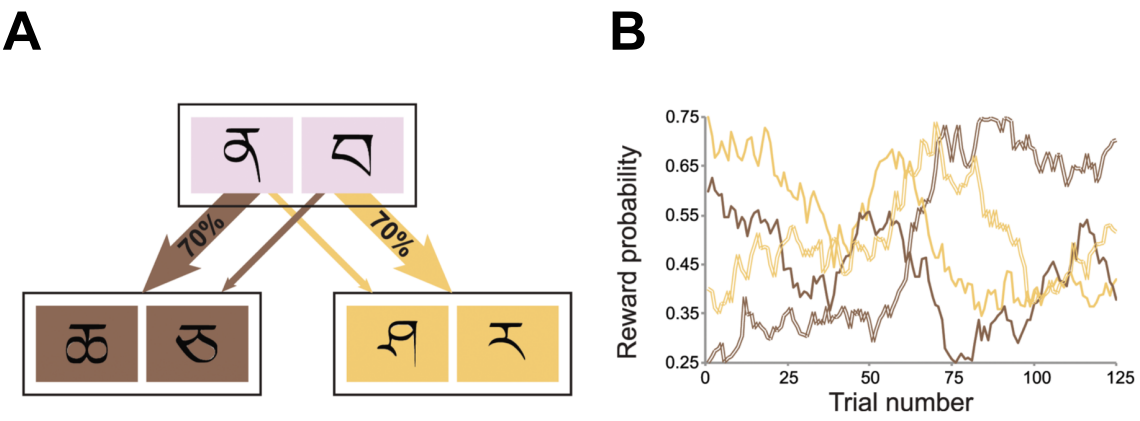

Figure 1. Design of the two-step task according to Daw et al. (2011). (A) Each choice in the first-step phase (pink) has a common (70% of trials) transition and a rare (30% of trials) transition to one of the two second-step phases (brown and yellow). (B) The reward probability of each choice in both second-step phases fluctuates gradually over time. Adapted from "When Does Model-Based Control Pay Off?" by W. Kool, F. A. Cushman, and S. J. Gershman, 2016, PLoS Computational Biology, 12(8), p. 3.

A Modified Two-Step Task

One of the motivations behind the two-step tasks is that more model-based learning leads to more accurate performance. However, when people have to make fast decisions, the flexible model-based system might be too slow to be more accurate (Heitz, 2014; Keramati, Dezfouli, & Piray, 2011). Additionally, people may consider cognitive demand as an effort cost and evaluate this effort cost in relation to accuracy benefits in decision making (Kool, McGuire, Rosen, & Botvinick, 2010). Therefore, it is optimal to use the model-free system when the accuracy does not increase with more cognitive effort or when the person needs to make a fast decision. Previous research showed that model-based strategies were not more accurate in the original

two-step task, meaning there is no trade-off between accuracy and demand (Akam, Costa, & Dayan, 2015; Kool, Cushman, & Gershman, 2016), hence people may not be sufficiently stimulated to use the model-based system. Kool et al. (2016) noted five features in the original two-step task that reduce the accuracy of the model-based system and introduced adjustments to the two-step task to solve these issues (we will refer to this as the modified two-step task).

The first feature in the original two-step task is that the probability of receiving a binary reward (i.e., chance of getting a reward or not) is not informative enough. One way to deal with this issue is to use a fluctuating number of points instead of fluctuating reward probabilities. In this case, each choice has a specific outcome (e.g., getting 4 points each time you choose that option), which is more informative (i.e., after every decision, the participant can observe the reward of their chosen option), and hence the accuracy of the model-based system increases. Secondly, the original two-step task uses rare transitions, meaning that the model-based choices sometimes lead to the non-preferred second-step phases. The authors showed that the relationship between the model-based strategy and accuracy increased when using deterministic transitions in the task (i.e., a choice of the first-step phase always leads to the same second-step phase). In third place, Kool et al. (2016) found that the difference between model-based and model-free strategies only affects the choices in the first-step phase. Thus, by removing the choices in the second-step phase, the importance of the first-step phase's choices increases, which in turn increases the accuracy-demand trade-off. Their fourth statement is that, in the original two-step task, the reward values always ranged between a lower and an upper bound that made the reward values too close to each other. Therefore, the outcomes may be too similar for both the model-based and model-free strategies. By increasing the range of the reward values, we can increase the differences between choices. And finally, the fifth factor is that the model-free strategy can adapt fast enough to the outcomes, to be as accurate as the model-based one, due to the slow changes in reward values. Using larger reward changes can solve this issue.

The Current Study

The current study used a task paradigm based on these changes proposed by Kool et al. (2016), so that our paradigm also includes an accuracy-demand trade-off. In our modified two-step task, participants had to obtain as many points (treasure coins) as possible. In each trial, participants were randomly presented with one of two possible first-step phases (i.e., two different states; Figure 2A). Each state of the first-step phase had a different pair of animals as stimuli, from which participants had to choose. Each animal deterministically led to a treasure chest (a second-step phase without choice options) that contained a specific number of points (reliable outcome; no reward probability).

Importantly, one of the animals in one state always led to the same number of points as one of the animals in the other state (Figure 2). The latter feature is of importance to dissociate between model-based and model-free strategies. For instance, a model-free learner will update the expected reward for each animal separately, while a model-based learner can also update the expected reward of the coupled animal in the other state as the model-based learner takes the structure of the task into account and, thus, generalizes the expected rewards over the two states.

Moreover, the outcome in our modified task could range between 0 and 10 points (Figure 2B). And unlike Kool et al. (2016), we used reward fluctuations of at least 3 points (large reward changes) instead of a Gaussian drift (i.e., random fluctuations over trials that vary according to a normal distribution), to maximize the competitive advantage of the model-based system. The current study used this modified two-step task to examine how participants behave in this two-step task and what makes a participant a better performer. We will elaborate on our research questions in the following sections.

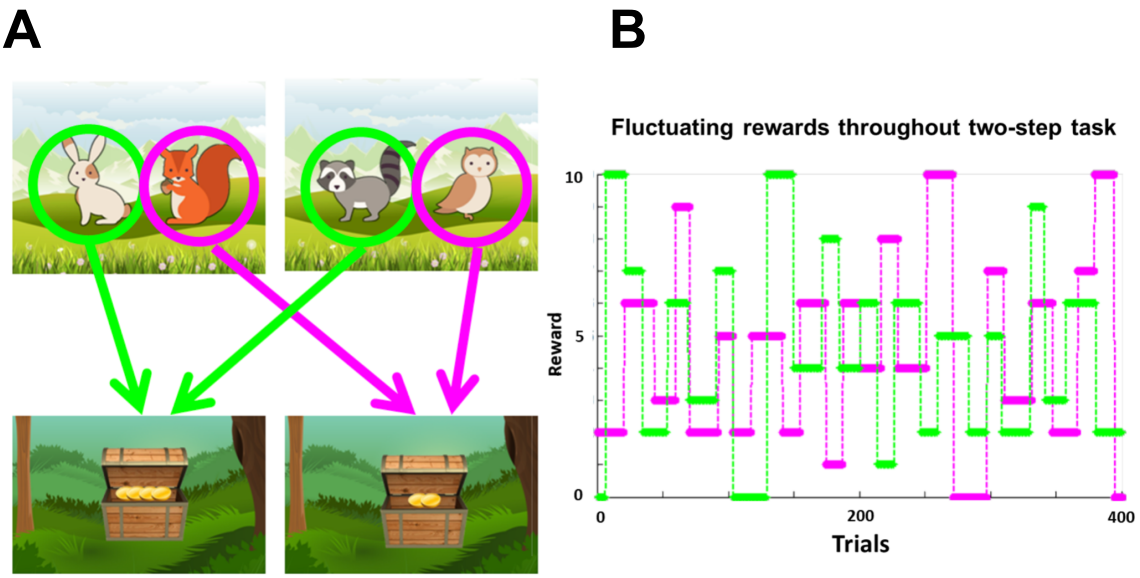

Figure 2. Task paradigm. (A) The two different states in which the participants had to select one of the animals. After choosing an animal, a treasure chest appeared, which the participants had to open to receive their reward (coins). A similar image was presented during the instructions of the task. (B) The reward fluctuations throughout the task. Each color represents the changing reward associated with one animal for each state.

How Model-Based Are Humans in the Two-Step Task?

To analyze how participants behave in the modified two-step task, we fitted computational models to a dataset and performed a model comparison analysis to see which computational model describes the behavior best. We had three main research questions for the model comparison. The first question was whether the participants used a mixture of model-based and model-free strategies, or mainly used model-based or model-free strategies in our task.

Daw et al. (2011) and subsequently many others (e.g., Kool et al., 2016; Piray et al., 2016) found that people use a mixture of model-based and model-free strategies in the two-step task. Recently, however, da Silva and Hare (2020) demonstrated that when the instructions before the two-step task explicitly explains all the features of the task, the use of model-based strategies is much higher than for the common instructions used in earlier studies. Moreover, da Silva and Hare (2020) found that some behavior in the two-step

task can be misinterpreted as model-free when it may actually be a model-based strategy using an incorrect model of the environment (e.g., wrong idea about the features of the task). In general, their findings suggest that participants mainly use model-based strategies in the two-step task.

In the current study, participants also received specific instructions about the task structure. This way, the participants in our task were also aware of the correct model of the task environment and, therefore, it is possible that participants in our task would show mostly model-based responses as well. Unlike the study by da Silva and Hare (2020), we used the modified two-step task. Kool et al. (2016) demonstrated that the modified task not only increases the benefits of a model-based learner, but that it also made participants rely more on model-based strategies. Therefore, we predicted that participants in our modified task will mainly use model-based strategies. To test this, we used a model-comparison approach to see if the data was better fitted by a hybrid model (i.e., computational model that still uses the balance parameter to arbitrate between model-based and model-free learning) or a pure model-based learner (i.e., computational model that only uses model-based learning). For completeness, we also compared the latter computational models with a pure model-free learner (i.e., computational model only using model-free learning).

The Exploration-Exploitation Trade-Off in the Two-Step Task

For our second research question, we focussed on a different dimension of learning and decision-making than the one between model-based and model-free learning, that might explain some of the behavior in the two-step task. As we previously noted, da Silva and Hare (2020) found that when the balance parameter of the two systems indicates that participants use less model-based strategies, the participants are not necessarily more model-free. Instead, their behavior might actually deviate from what is expected from a model-based learner due to the use of an incorrect model of the environment. Furthermore, Akam et al. (2015) demonstrated that model-free responses can be misclassified as being model-based due to exploiting a regularity in the task structure. In other words, certain behaviors in the two-step task are not determined by the dichotomy between model-based and model-free systems

(Collins & Cockburn, 2020; Daw, 2018) but can be misclassified as such, which limits our understanding of what actually happens during this task. Therefore, we decided to also look at a different dimension than the one between model-based and model-free learning. Particularly, we focussed on a dimension involved in reward-based decision-making, namingly exploration-exploitation trade-off.

In reinforcement learning, the parameter *inverse temperature* is typically introduced to represent the exploration-exploitation trade-off (Sutton & Barto, 2018). Exploitation refers to an agent selecting a choice that (s)he considers to be the most optimal choice, while exploration refers to selecting another choice because the agent might want to learn if the other option is better or not. Therefore, finding a balance between exploration and exploitation is important for making optimal decisions, because too much exploitation makes it impossible to learn the true outcome of other options and too much exploration makes decisions too random (i.e., less goal-directed). So, exploration is more important in unknown tasks, while exploitation is better when the environment is already known.

The trade-off between exploration and exploitation is especially important for adapting to changing environments, such as outcomes changing over time. However, whether someone will explore or exploit depends on many different factors (Cohen, McClure, & Yu, 2007). One possible factor that may affect this trade-off is the belief that an outcome change occurs more often when something else changes in the environment. Therefore, we were interested in whether an environmental change in our task, other than an outcome change, prompts people to explore or exploit more.

In our task, the outcomes of the choices fluctuated over time, so it was preferable for the subjects to explore from time to time to see whether or not the outcome of the other choice had changed to a better one. Another feature in our task, which was particularly important for model-based learning, was that there were two different environments with different stimuli (herein called states) alternating over time; the stimuli could either stay the same or change to different stimuli as in the previous trial. However, it is possible that people

associate these changing stimuli with outcome changes, and that they base their decisions on the incorrect assumption that outcome changes accompany stimuli changes. Thus, the current study investigated if participants used a different exploration-exploitation trade-off when the state changed than when the state stayed the same. We predicted here that participants explore more after a state change because a state change might evoke a belief of an outcome change.

High Learning Rates for Reliable Outcomes

Our third research question was how fast participants learn the outcomes associated with the choices in our two-step task. We predicted that participants would immediately learn the correct outcome associated with a decision in our task, because our task used completely reliable outcomes instead of the outcome probabilities used in the original two-step task; hence it is optimal to have a high learning rate. In reinforcement learning, a learning rate parameter captures the rate at which participants learn the value of a choice based on the previous perceived outcome (Sutton & Barto, 2018). Thus, we examined if this learning rate parameter in our task is better fixed on its maximum value or not.

Relationships Between Parameters and Accuracy

Furthermore, we had two more research questions involving the relationships between individual differences in behavior and optimal decision-making. Kool et al. (2016) modified the two-step task so that model-based control would increase accuracy and demonstrated that this was the case in the modified task. Therefore, we were interested in how a better performer actually behaves in our two-step task. As our task was also modified to maximize the benefits of model-based control, our fourth research question was whether individuals with higher model-based control would also have higher accuracy scores in our task.

Our fifth research question was how the trade-off between exploration and exploitation is related to optimal decision-making in our task, because the exploration-exploitation trade-off plays an important role in selecting optimal decisions and some of its tendencies to explore or exploit can be suboptimal (Sutton & Barto, 2018). Therefore, it might be interesting to see if the tendency

to explore or exploit is simply in general correlated to accuracy, or whether instead the exploration-exploitation trade-offs when a state changed or stayed the same have a different relationship with accuracy. These correlations might tell us something about how more optimal decision-makers behave in the two-step task.

Method

Participants

Our sample size consisted of 275 college students performing a two-step task. This dataset was collected by Prof. Dr. Qi Chen and Prof. Dr. Tingyong Feng at Southwest University in Chongqing, China. In total, twelve participants were excluded from the data analysis. Two subjects were excluded because they did not complete the whole experiment, resulting in a large amount of missing data for these subjects. The other ten participants were excluded because their accuracy scores (i.e., number of trials choosing the optimal choice) were more than 2 standard deviations below the mean (i.e., scores below 61% accuracy). This leaves us with a total sample size of 263 participants (189 females, 74 males; mean age: 20.07; range: 17–30 years of age). Each participant received 25 yuan (approximately €3.25) for their participation. The data exclusion and the analysis of demographic information was carried out using R (version 4.0.5)[1].

Behavioral Task

The participants performed a two-step task on a computer in a laboratory. The goal of the two-step task was to obtain as many points (treasure coins) as possible. This task had two states of the first-step phase (Figure 2A), each with two different animals of which the participants had to choose one on every trial. The participants could select an animal by pressing the “F” or “J” button on the keyboard of a computer. The time limit of the response was 2500 ms. Immediately after choosing an animal, a treasure chest appeared representing the second-step phase. To obtain the reward of the selected stimulus in a state, the participants were instructed to press the spacebar on the keyboard when the treasure chest was presented. Here, the time limit of responding was 3500 ms. Each chosen animal deterministically led to an outcome within an outcome range of 0 to 10 points.

Importantly, the reward of each animal in a state is coupled to the reward of one of the animals in the other state (Figure 2). For example, stimulus A of one state and stimulus C of the other state always had the same reward, while

[1] All scripts for the data analysis can be found on the following link: https://github.ugent.be/fdspiege/Thesis_2021.git

stimulus B had the same reward as stimulus D. This is an important feature of the task that allows us to dissociate between the use of model-free and model-based strategies. For instance, the model-based system uses the structure of the task (i.e., the coupling between pairs of stimuli in different states) by also updating the expected reward of the other animal after seeing the shared reward of an animal in the other state. In contrast, the model-free system only uses the experience of a received reward, therefore only updating the expected reward of the chosen animal and not generalizing the expected reward to the linked animal in the other state. Thus, the model-based system updates the expected reward of the stimuli more efficiently, resulting in a higher total score.

To encourage learning throughout the task, the corresponding reward of each choice changed eight times per 100 trials with a minimal reward shift of 3 points and had a standard deviation of at least 3 (Figure 2B). These reward changes were statistically independent of each other, except that the reward transitions could never occur simultaneously and the choices in a state never had the same rewards (so there always was an optimal choice). Moreover, the reward fluctuations for stimulus A and C in the first half of the experiment were mirrored in time and assigned to stimulus B and D in the second half of the experiment, and vice versa. This ensured that the stimuli could total the same number of points throughout the task, while still keeping the predictability of the rewards minimal. In addition, the reward fluctuations in the task were generated independently for each participant.

Before the experiment started, the participants performed a practice block of ten trials. Then an instruction screen appeared explicitly explaining that there are two treasure chests, of which the amount of treasure changes throughout the task, but each animal always goes to the same treasure chest as one of the other animals. After the instructions, the participants performed another practice block of ten trials. The practice blocks had different stimuli (i.e., different animals) than in the experimental block. This training phase allows participants to learn about different aspects of the task (such as range of rewards, reward changes, and shared rewards between stimuli in different

states) without generalizing the expected rewards of these stimuli to the stimuli in the experimental phase. The experimental phase consisted of 400 trials in total.

Computational Models

First of all, we used a Rescorla-Wagner update rule (Rescorla & Wagner, 1972) to capture the model-free and model-based learner. The model-free learner updates the expected reward V_{MF} of a specific stimulus (A, B, C, or D) based on the difference between the previous expected reward and the new obtained outcome of that stimulus. For example, the model-free learner chooses animal A, which the model-free learner predicts to have a reward of four coins based on the previous trials. After selecting animal A, a reward of (say) ten coins is obtained. By subtracting the previous expected reward (4 points) from the new outcome (10 points) we get a discrepancy of six which is called the reward prediction error. Additionally, there is also a learning rate parameter α which has a value between 0 and 1. This parameter determines how fast the algorithm will learn. If the learning rate parameter is equal to 0, it means that the agent is not learning anything about the reward of a stimulus. If the learning rate is equal to 1, the agent completely changes the expected reward of the stimulus into the last obtained outcome, O, of the stimulus itself. The reward prediction error can be used to compute the new expected reward, $V_{MF}(A_t)$, by first multiplying it with the learning rate parameter, which we then add up with the previous expected reward, $V_{MF}(A_{t-1})$. The expected reward of a model-free learner can be computed as follows:

$$V_{MF}(A_t) = V_{MF}(A_{t-1}) + \alpha * (O - V_{MF}(A_{t-1})) \quad (1)$$

As previously explained, the model-free learner updates the expected reward for each stimulus separately. In contrast, the model-based learner can also update the expected reward of the coupled animal from the other state. For example, the expected reward of animal A, $V_{MB}(A_t)$, can be generalized to

animal C, $V_{MB}(C_t)$. Thus, the expected reward of a model-based learner can be computed as follows:

$$V_{MB}(A_t) = V_{MB}(A_{t-1}) + \alpha * (O - V_{MB}(A_{t-1})) \quad (2)$$

and

$$V_{MB}(C_t) = V_{MB}(A_t) \quad (3)$$

We assume that these expected rewards are independently computed and are combined to arrive at a mixed value, upon which the behavior is based. This was done by a hybrid account which computes a parameter (w) that captures the balance between the model-based and model-free learner, in which the expected rewards of the two systems are weighted by the parameter for each stimulus. By capturing the balance between the two learning systems, the expected reward of an agent can be computed as follows:

$$V(A_t) = w * V_{MB}(A_t) + (1 - w) * V_{MF}(A_t) \quad (4)$$

where w is also bounded between 0 and 1. When $w = 0$, we have a pure model-free learner. When $w = 1$, we have a pure model-based learner. Thus, a higher value of w means we have a stronger reliance on a model-based system, whereas a lower value is associated with a higher degree of model-free strategies. Note here that the expected rewards for each stimulus had an initial value of 5 (i.e., in the beginning of the task), as the rewards could range between 0 and 10.

Furthermore, we compute the decisions by using a softmax rule that translates the expected rewards of an agent to probability of choice (e.g., probability of picking animal A instead of animal B):

$$P_t(A) = \frac{exp(\beta * V(A_t))}{exp(\beta * V(A_t)) + exp(\beta * V(B_t))} \quad (5)$$

where β is the inverse temperature parameter that reflects the exploration-exploitation trade-off. When $\beta \to 0$, the decisions tend toward completely random (i.e., exploratory) choices. When $\beta \to \infty$, the decisions tend toward fully exploiting one's current knowledge, and thus toward choosing the stimulus with the highest expected reward.

Furthermore, people sometimes tend to repeat their responses (e.g., perseveration of the same button press) or repeat the same choices (e.g., choosing the same stimulus) regardless of the reward outcomes. Kool et al. (2016) introduced two "stickiness" parameters in the softmax decision rule, which represent an agent's tendency to repeat the same choices (referred to as *stimulus stickiness*) and to repeat the same responses (referred to as *response stickiness*) independent of expected rewards. Based on Kool et al. (2016), we also introduced these parameters in our models, because they may capture some of the tendencies in our task that were not reward-based. In this case, the softmax decision rule can be computed as follows:

$$P_t(A) = \frac{exp(\beta^*[V(A_t) + \pi^*rep(A) + \rho^*resp(A)])}{exp(\beta^*[V(A_t) + \pi^*rep(A) + \rho^*resp(A)]) + exp(\beta^*[V(B_t) + \pi^*rep(B) + \rho^*resp(B)])} \quad (6)$$

where the variable $rep(A)$ represents whether (= 1) or not (= 0) the stimulus A is the same as the previously chosen stimulus. This variable $rep(A)$ is weighted to the parameter stimulus stickiness π, which captures the extent to which the agent chooses the same stimulus or not between two consecutive trials. This parameter is added to the expected reward $V(A_t)$. When $\pi > 0$, the agent is more likely to choose the same stimulus. When $\pi < 0$, the agent is more likely to switch choices between two trials. The variable $resp(A)$ represents whether (= 1) or not (= 0) the position of stimulus A (i.e., left or right) requires the same response as the previously made button press. Here, $resp(A)$ is weighted by response stickiness ρ, which captures the extent to which the agent presses the same button or not, and is then also added to the expected reward $V(A_t)$. When $\rho > 0$, the agent is more likely to respond with the same button press. When $\rho < 0$, the agent is more likely to switch button presses. In other words, by including these two stickiness parameters, the probability of choosing stimulus

A also depends on whether or not stimulus A was selected in the previous trial and selecting this stimulus requires the same button press previously made or not.

As we explained in the introduction, we were interested in whether the trade-off between exploration and exploitation differed when a state changed between trials compared to this trade-off when the state remained the same as in the previous trial. For this, we also introduced separate inverse temperatures in the softmax decision rule: one for when the state changes (i.e., different animals than in the previous trial), β_{Change}, and another for when the state is the same as in the previous trial, β_{Same}. We will refer to these two inverse temperatures as the dual inverse temperatures. Note that the trials with different states for inverse temperature β_{Change} also includes the first trial of the experiment, assuming the state has changed even though there was no previous trial. Additionally, we first checked if the number of trials for each inverse temperature were relatively similar. Even though a paired samples t-test revealed that the number of trials for β_{Change} (M = 203.24, SD = 9.50) were significantly higher than the number of trials for β_{Same} (M = 195.18, SD = 9.51), $t(262) = 6.96$, $p < 0.001$, this difference was only small: about eight trials out of the 400. Therefore, the dual inverse temperatures can be estimated based on a relatively similar number of trials.

Model Fitting and Model Selection

All trials on which participants reached the time limit to respond (i.e., not selecting a stimulus or not opening the chest) were first excluded (an average of 0.4% of all trials). No participants were excluded from the data analysis due to their number of missed responses, as the maximum number of missed responses was 25 trials (6.25% of 400 trials). For the model fitting, we used the *mfit* toolbox (https://github.com/sjgershm/mfit; Gershman, 2016) running on MATLAB, version R2020b (The Math Works, Inc., http://www.mathworks.com/).

We estimated the free parameters of each model for each subject separately to maximize the log-likelihood of the data. To estimate the parameters, we used the maximum *a posteriori* (MAP) estimate (i.e., the mode

of the posterior distribution) with empirical priors (i.e., prior distributions estimated from another dataset), based on Gershman (2016). Using prior distributions can help with the identifiability of models, which is a critical factor when parameter estimates need to be interpreted. Models are unidentifiable when different combinations of parameter settings yield the same likelihoods (e.g., different settings of learning rate and inverse temperature having equivalent likelihoods). Specifically, we used a gamma distribution (i.e., a continuous probability distribution with a shape parameter and scale parameter), $\beta \sim \text{Gamma}(4.83, 0.73)$, as prior distribution of the inverse temperature when no stickiness parameters were introduced and, $\beta \sim \text{Gamma}(4.82, 0.88)$, when at least one of the stickiness parameters was included. The inverse temperatures always had lower and upper bounds of 0 and 50, respectively. The stickiness parameters had prior distributions, $\pi, \rho \sim \mathcal{N}(0.15, 1.42)$, with boundaries -5 and 5. Both learning rate parameter α and balance parameter w had uniform priors, i.e. prior distributions of equal probability for any value between the boundaries 0 and 1. For each participant, we used random initializations for the optimization of all parameters, and repeated this optimization process ten times. We then selected the parameter estimates from the optimization with the maximum log-likelihood of these ten iterations.

For the model comparison, we used both Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to select the best-fitting models. AIC and BIC are both based on the likelihood, but include a penalty term for the number of model parameters. BIC has a stronger penalty term (i.e., is more conservative) than AIC. These penalty terms are used to avoid overfitting. The log-likelihood is by definition at least as good (i.e., higher log-likelihoods) for nested models with more parameters, which are not necessarily better models as they may be worse in predicting new datasets. Therefore, we decided to use AIC and BIC. Here, the best model is considered to be the one with the lowest AIC or BIC, indicating there is a good balance between a high likelihood and a low complexity.

First, we used AIC and BIC to compare models with or without the stickiness parameters π and ρ (i.e., both, one or none of these parameters), to see if we would include these parameters in the models in further model comparisons. We did this based on the method of Kool et al. (2016) and because the stickiness parameters were not the main focus of our research questions.

Secondly, we compared models using AIC and BIC based on three different aspects. The first aspect we wanted to look at was whether the participants were fully model-based, fully model-free, or instead used a mixture of the two systems (i.e., hybrid model). In this case, we compared models in which parameter w was fixed to 1, fixed to 0, or a free parameter. Here, we predicted that a pure model-based model would be a better fit to our data, because the instructions in our task also explicitly explained the features of the task as in the study by Silva and Hare (2020). As a second aspect, we also compared models to examine whether the inverse temperature is better off being split up into β_{Change} and β_{Same}; one when the state changed and one when the state stayed the same. Here, we would like to note again that when a state changed, simply different stimuli were presented in that trial than in the previous trial, and that these state changes were independent of the reward changes. Thus, the participants had to map the expected rewards of the stimuli in one state to the other when a state changed. By introducing these dual inverse temperatures, we investigated if a different exploration-exploitation trade-off was used when the state changed or stayed the same. It is possible that the changing states have an effect on participants' beliefs about the reward changes. Thus, we predicted that participants explore more when a state changes than when it stays the same, as a changing state may elicit higher belief of reward changes. The third aspect we looked at was whether or not the learning rate α should be fixed to 1. In our task, the reward outcome of a stimulus was completely reliable; choosing a stimulus deterministically led to a specific reward (e.g., stimulus A has a reward of 7 points). Therefore, participants should be able to immediately learn the correct value of the reward that they perceived (i.e., a high learning rate).

Relations Between Parameters and Accuracy

We also examined individual differences between the participants, using Python, version 3.7.10 (Python Software Foundation, https://www.python.org/). As model-based learners are able to use environmental information more efficiently to learn about the rewards of stimuli (e.g., $V_{MB}(C_t) = V_{MB}(A_t)$), model-based learners may in general be more accurate in making decisions than model-free learners. However, previous research demonstrated that model-based learners in the original two-step task do not actually have a higher accuracy rate (Akam et al., 2015; Kool et al., 2016). To remedy this, Kool et al. (2016) adjusted the two-step task so that people who used a model-based strategy, would also have accuracy. The task in our study was based on the adjustments created by Kool et al. (2016). Unlike Kool et al. (2016), we also worked with large reward fluctuations (i.e., minimum reward shifts of 3 points) rather than a Gaussian drift (i.e., random fluctuations according to a normal distribution), to maximize the competitive advantage of the model-based system. Therefore, we also investigated whether participants who used more model-based strategies would have higher accuracy scores in our task. For this purpose, we correlated the balance parameter w with accuracy. For the accuracy scores, we specifically looked at the percentages of trials in which optimal decisions were made (i.e., choosing the animals with the highest rewards) and not the total points obtained. We chose the amount of optimal decisions because the possible number of points that could be earned were randomly assigned to participants and did not particularly represent the accuracy of a participant.

In addition, we were also interested in the relationship between exploration-exploitation trade-off and accuracy, because this trade-off is important for making optimal decisions (Sutton & Barto, 2018). Here, we were particularly interested in whether there is simply a general relationship between the tendency to explore or exploit and optimal decision-making, or whether instead the exploration-exploitation trade-offs in case a state changed or stayed the same had a different relationship with accuracy. Examining the latter might tell us something about how more optimal decision-makers behave in our task.

We first correlated the individual inverse temperatures β with the accuracy scores, to see if the exploratory tendency of participants are in general related to the amount of making optimal decisions. Then, we tested whether the dual inverse temperatures β_{Change} and β_{Same} had a different relationship with accuracy scores. Here, we correlated accuracy with β_{Change}, β_{Same}, the mean of the two inverse temperatures (i.e., $(\beta_{Change} + \beta_{Same})/2$), and differences between the two inverse temperatures (i.e., $\Delta\beta = \beta_{Change} - \beta_{Same}$). In our task, the stimuli deterministically led to a specific number of points and the reward changes occured only eight times per 100 trials for each choice. Moreover, there were only two options to choose from. So, exploring the other choice, instead of the one expected to be the optimal choice, can more often lead to suboptimal decisions. Therefore, we predicted that more exploitation (higher inverse temperatures) in general would lead to more optimal decisions (higher accuracy scores).

Results

Model Fitting and Model Selection

Based on Kool et al. (2016), we first tested whether models including response and stimulus stickiness generated a better fit. Here, both AIC and BIC were compared among four models; models with both stickiness parameters, one of either, or none. This already allows for including or excluding these stickiness parameters in further model comparisons. AIC and BIC both penalize for the number of parameters, the difference is that AIC is less conservative. According to both AIC and BIC, the model with both stickiness parameters was a better fit to the data (i.e., lowest values for AIC, BIC, and the negative log-likelihood; Table 1).

Table 1

Comparing models for inclusion of response and stimulus stickiness parameters.

Model	Parameters	BIC	AIC	- LL	% best BIC	% best AIC	% best LL
Two stickiness parameters	w, α, β, π, ρ	**41598**	**36355**	16862	**100%**	**100%**	100%
No response stickiness	w, α, β, π	67738	64543	30719	0%	0%	0%
No stimulus stickiness	w, α, β, ρ	49822	45627	21762	0%	0%	0%
No stickiness parameters	w, α, β	84375	81229	39825	0%	0%	0%

Note. The last three columns represent how many participants on an individual level had a better BIC, AIC, or LL (log-likelihood) for each model.

Subsequently, we compared the other models by already including both stickiness parameters. Here, we analysed three different aspects. The first one is whether the inclusion of a free parameter w (i.e., hybrid model) or fixing w to either 1 (i.e., pure model-based) or 0 (i.e., pure model-free) was a better fit. Secondly, we investigated if there is a difference in the exploration-exploitation trade-off between states that stayed the same and states that changed. To do so, we also introduced dual inverse temperature parameters for different or

same states; one for trials in which the previous stimuli were different stimuli than the current one (β_{Change}) and one for trials in which the same stimuli were repeated (β_{Same}). And finally, we also tested whether the learning rate α is better fixed to 1. We tested this because a stimulus in our task deterministically led to a specific number of points, which makes the stimulus outcome completely reliable, and it is better to have a high learning rate when a reward is completely reliable.

Here, we also compared both AIC and BIC of each model. According to the AIC, the best fit was the pure model-based model with a fixed learning rate to 1 and dual inverse temperatures (Table 2). Although a large number of the individual AICs (30%) fitted better with the pure model-based model with fixed learning rate to 1, most of the individual AICs (42.2%) were better for this latter model with dual inverse temperatures. Based on the BIC, a pure model-based model with a fixed learning rate to 1 was the best fit. For most participants (71.5%), the individual BIC was at lowest (best) for this pure model-based model with fixed learning rate to 1. However, in almost a quarter of the BIC cases (23.2%), the better fit was for the pure model-based model with fixed learning rate to 1 and dual inverse temperatures. These results indicate that, in our two-step task, the participants mainly used model-based strategies and quickly learned the correct reward values of the stimuli after obtaining them. Moreover, the inverse temperature is better off being split up into β_{Change} and β_{Same} according to the AIC. Therefore, we decided to further explore the dual inverse temperatures to see if different exploration-exploitation trade-offs were used when a state changed or not.

Table 2

Comparing the hybrid, pure model-based, and pure model-free models with or without a fixed learning rate to 1 and dual inverse temperatures.

Model	Parameters	BIC	AIC	- LL	% best BIC	% best AIC	% best LL
Hybrid	w, α, β, π, ρ	41598	36355	16862	0%	0%	4.2%
Hybrid + fixed LR (α = 1)	w, β, π, ρ	40161	35966	16931	1.1%	6.5%	0.8%
Hybrid + dual IT	w, α, β_1, β_2, π, ρ	42156	35864	16354	0%	0%	40.7%
Hybrid + fixed LR (α = 1) & dual IT	w, β_1, β_2, π, ρ	40735	35492	16431	0%	5.7%	23.6%
MB (w = 1)	α, β, π, ρ	40158	35964	16930	1.1%	5.3%	3%
MB (w = 1) + fixed LR (α = 1)	β, π, ρ	**38823**	**35677**	17049	**71.5%**	**30%**	1.1%
MB (w = 1) + dual IT	α, β_1, β_2, π, ρ	40779	35535	16453	0%	8.4%	19%
MB (w = 1) + fixed LR (α = 1) & dual IT	β_1, β_2, π, ρ	**39417**	**35223**	16559	**23.2%**	**42.2%**	7.2%
MF (w = 0)	α, β, π, ρ	46414	42220	20058	0%	0%	0.4%
MF (w = 0) + fixed LR (α = 1)	β, π, ρ	44937	41791	20106	3%	1.1%	0%
MF (w = 0) + dual IT	α, β_1, β_2, π, ρ	47799	42556	19963	0%	0%	0%
MF (w = 0) + fixed LR (α = 1) & dual IT	β_1, β_2, π, ρ	46302	42108	20002	0%	0.8%	0%

Note. The last three columns represent how many participants on an individual level had a better BIC, AIC, or LL (log-likelihood) for each model. MB stands for pure model-based, MF for pure model-free, LR for learning rate, and IT for inverse temperature. In the parameters column, β_1 and β_2 represent β_{Change} and β_{Same}.

To further test if the tendency to explore changed based on whether a state changed or not, we examined whether the dual inverse temperatures were significantly different from each other. For this analysis, we selected the model which had the best fit according to the AIC. Here, we first obtained the average estimate of each parameter and the average standard error (SE) of each parameter. To calculate the SEs of the parameters, we used the second

derivatives of each likelihood based on Hessian matrices (also returned by the *mfit* toolbox in MATLAB) and measured the square root of the diagonals of minus the inverse of the Hessian matrices (Verguts & Storms, 2004). Note in Table 3 that the estimates of all parameters, in the model with the best AIC, are relatively precise (i.e., relatively low SEs).

Table 3

The mean estimates and SEs of each parameter for each model interpreted.

	Model with best AIC		Model without stimulus stickiness		Model with parameter w		Model with best BIC	
Parameters	Mean	SE	Mean	SE	Mean	SE	Mean	SE
β_{Change}	4.53	1.81	4.57	0.82	-	-	-	-
β_{Same}	2.88	1.77	5.66	1.07	-	-	-	-
ρ	1.79	1.38	1.58	0.59	1.68	1.24	1.71	2.53
π	1.72	1.35	-	-	1.36	1.56	1.39	2.85
β	-	-	-	-	4.06	1.95	3.94	3.08
w	-	-	-	-	0.87	0.33	-	-

Note. The SEs of the parameters are based on the Hessian matrices.

Subsequently, we tested if the normality assumption holds for the two inverse temperature parameters β_{Change} and β_{Same}, to choose whether to use a parametric or nonparametric test. A Shapiro-Wilk test (Shapiro & Wilk, 1965) revealed a non-normal distribution for both parameters β_{Change}, $W(263) = .94$, $p < .001$, and β_{Same}, $W(263) = .95$, $p < .001$ (Figure 3). Therefore, we decided to perform a nonparametric Wilcoxon signed-rank test (Wilcoxon, 1945) to see if the dual inverse temperature parameters were significantly different from each other.

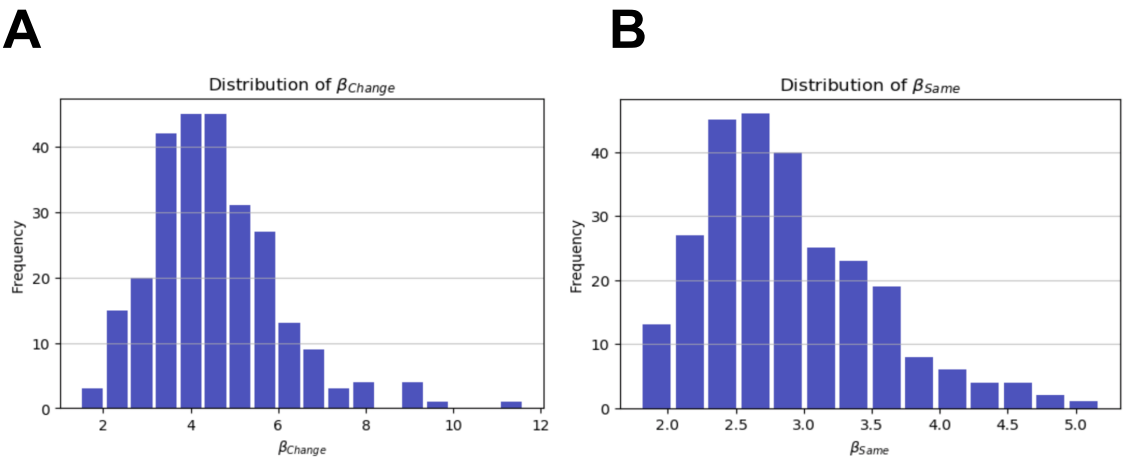

Figure 3. Frequency distributions of parameters β_{Change} (A) and β_{Same} (B).

The Wilcoxon signed-rank test revealed that parameter β_{Change} (*Mdn* = 4.41) was significantly higher than parameter β_{Same} (*Mdn* = 2.76), *z* = 1,246.00, *p* < .001. This would suggest that the participants exploited more after a state changed than when a state stayed the same as in the previous trial. This is inconsistent with our prediction that exploration should be higher when a state changes, because it may evoke the idea that a reward value also changed.

However, it is reasonable to think that the stimulus stickiness parameter π might influence this result. Repeating the same choice and exploitation show similar behavioral responses (i.e., they both represent choosing the same stimulus), and choice repetition is only possible when the same state occured, because only then can the same stimulus be selected. So, the stimulus stickiness parameter can only apply to the same states. Therefore, it is possible that the stimulus stickiness π specifically interfered with inverse temperature β_{Same}, and resulted in β_{Same} being significantly lower than β_{Change} in this model. In other words, the stimulus stickiness parameter may have partially absorbed exploitation from the inverse temperature in the same states. Therefore, we decided to also perform a *post-hoc* analysis, exploring whether a similar difference between the dual inverse temperatures appears when the stimulus stickiness parameter is excluded; a model with dual inverse temperatures and only response stickiness. We see for this model without stimulus stickiness that

the estimates of each parameter are relatively accurate (i.e., relatively low SEs; Table 3). However, it is important to note here that this model without stimulus stickiness yields a worse fit based on both BIC (a goodness-of-fit estimation of 48,146) and AIC (a goodness-of-fit estimation of 45,000) than almost every model we compared.

Also here, we decided to use a non-parametric test, because a Shapiro-Wilk test revealed a non-normal distribution for both parameters β_{Change}, $W(263) = .95$, $p < .001$, and β_{Same}, $W(263) = .97$, $p < .001$ (Figure 4). Interestingly, a Wilcoxon signed-rank test revealed that parameter β_{Change} (*Mdn* = 4.41) was significantly lower than parameter β_{Same} (*Mdn* = 5.57), $z = 4{,}815.00$, $p < .001$, in the model without stimulus stickiness. This is consistent with our predictions and suggests that the stimulus stickiness π interfered with the inverse temperature parameter β_{Change}.

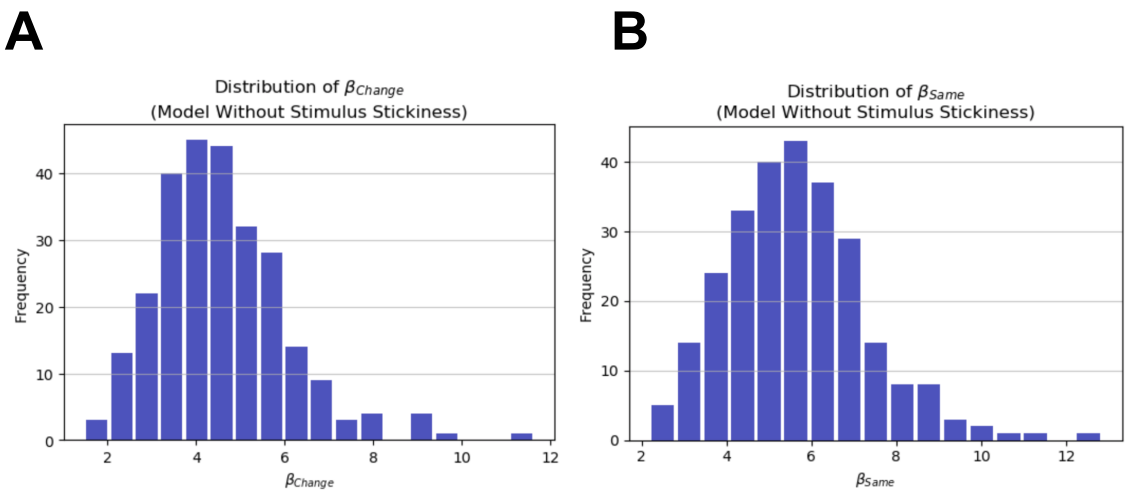

Figure 4. Frequency distributions of parameters β_{Change} (A) and β_{Same} (B) in the model without the stimulus stickiness parameter.

One possible explanation for this effect of state transition (i.e., changed stimuli) on exploration could be that the participants believed a reward is more likely to change after a state changed than when the state stayed the same.

Therefore, we performed a *post-hoc* analysis to test whether it was actually the case that there were more reward changes when the state changed. Here, we also decided to use a Wilcoxon signed-rank test, as a Shapiro-Wilk test revealed a non-normal distribution for both the percentages of trials having reward changes after a changed state, $W(263) = .98$, $p < .001$, and reward changes after a unchanged state, $W(263) = .98$, $p = .004$. The results of a Wilcoxon signed-rank test showed no significant difference between percentages of trials having reward changes after a changed state (*Mdn* = 11.70) and reward changes after unchanged states (*Mdn* = 11.06), $z = 15{,}487.00$, $p = .130$. So, if participants would assume more reward changes occured after state changes, this would be a false assumption (i.e., incorrect model of the task). Our data did not allow us to test this explanation any further.

Relations Between Parameters and Accuracy

Although both AIC and BIC prefered pure model-based models, we were still interested in whether participants who used more model-based strategies selected more optimal choices in our task. Therefore, we also correlated parameter w and accuracy scores (i.e., % of trials making optimal decisions). For this analysis, we used the parameters w of the hybrid model with the best BIC (i.e., hybrid model with fixed learning rate to 1), as BIC is more conservative. For this model including a free parameter w, we can see that the estimates of most parameters are relatively precise (i.e., relatively low SEs), except for parameter w as its estimates can only range between 0 and 1 (Table 3).

We first tested whether the normality assumption holds to perform a parametric correlation between parameter w and accuracy score. A Shapiro-Wilk test demonstrated a significant deviation from normality for both parameter w, $W(263) = .79$, $p < .001$, and accuracy score, $W(263) = .99$, $p = .025$ (Figure 5). Parameter w had a skewness of -1.62 (*SE* = 0.15) and a kurtosis of 2.67 (*SE* = 0.3), and accuracy score had a skewness of -0.34 (*SE* = 0.15) and a kurtosis of -0.11 (*SE* = 0.3). In Figure 5A, it is clear that parameter w is negatively skewed, likely due to a ceiling effect. Therefore, we decided to perform a Spearman (nonparametric) correlation.

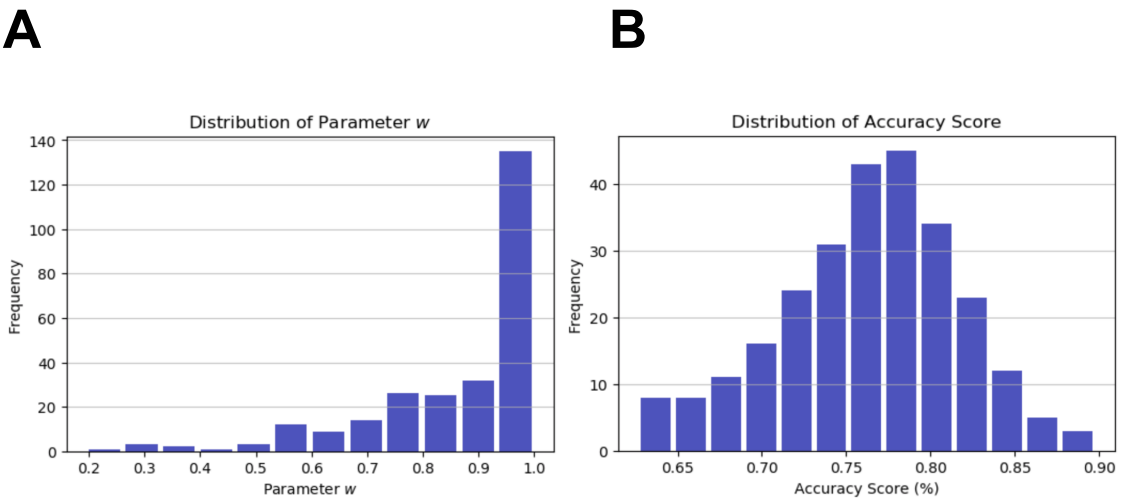

Figure 5. Frequency distributions of parameter w (A) and accuracy score (B).

The Spearman correlation reported no significant correlation between parameter w and accuracy scores, $r(263) = .09$, $p = .157$ (Figure 6). Thus, we can conclude that using more model-based strategies in our task is probably not related to higher accuracy scores. However, we should be cautious interpreting these results as they may be due to a ceiling effect.

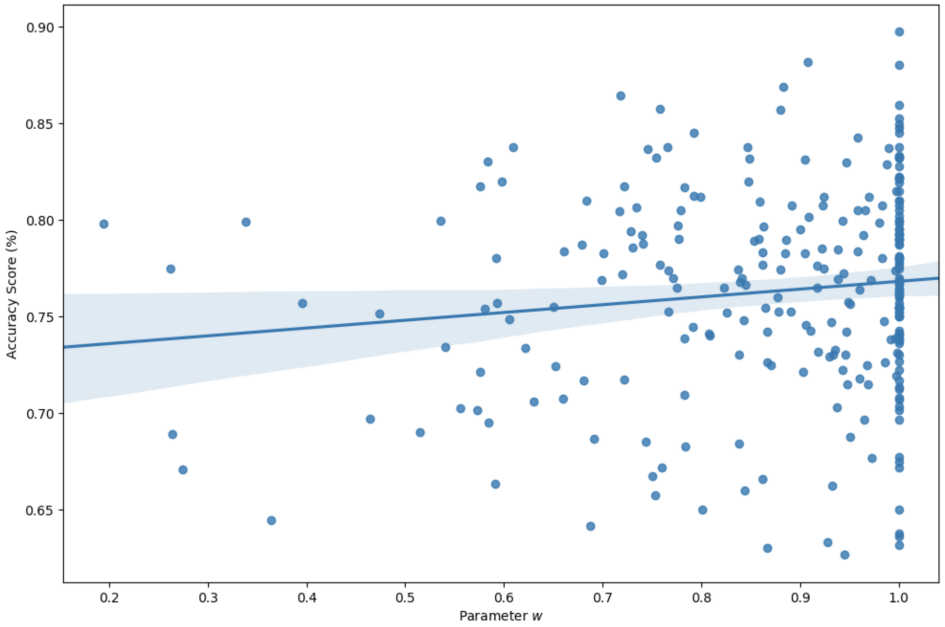

Figure 6. Scatterplot showing the relationship between accuracy scores and balance parameter w. The shading represents the 95% confidence interval.

In addition, we wanted to see if exploratory tendencies were related to making optimal decisions. To do so, we first correlated the inverse temperature β (of the model with the best fit based on BIC) with accuracy score. For the model with the best BIC, we see that the estimates of inverse temperature β are relatively accurate (i.e., relatively low SEs), but not for the stickiness parameters π and ρ as their values can only vary between -5 and 5 (Table 3). We first performed a Shapiro-Wilk test that showed a non-normal distribution for parameter β, $W(263) = .94$, $p < .001$. Subsequently, a Spearman correlation reported a significant positive correlation between parameter β and accuracy, $r(263) = .35$, $p < .001$ (Figure 7). This suggests that participants who exploit more during the task make more often optimal decisions in our task.

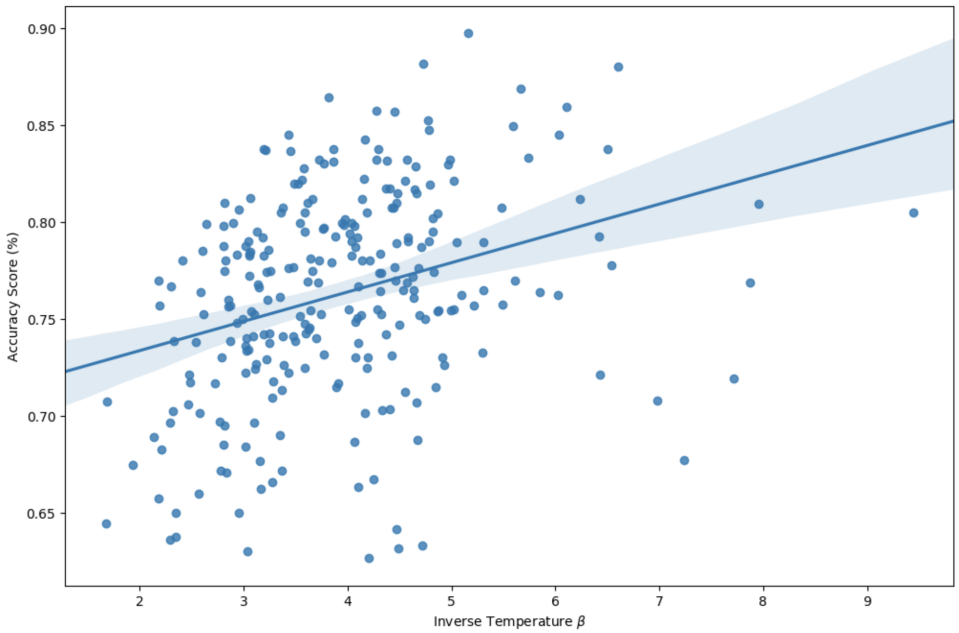

Figure 7. Scatterplot showing the relationship between accuracy scores and inverse temperature β (in the model with best fit according to BIC). The shading represents the 95% confidence interval.

Furthermore, we were interested in whether the two exploration-exploitation strategies used, when the state changed and when the

state stayed the same, were differentially related to how optimal decisions were. To do so, we correlated the two inverse temperatures β_{Change} and β_{Same} with accuracy. In addition, we also correlated both the difference between the dual inverse temperatures, $\Delta\beta$ ($\beta_{Change} - \beta_{Same}$), and the mean of these two parameters with accuracy. For these correlations, we selected the dual inverse temperatures of the model which had the best fit according to AIC. A Spearman (nonparametric) correlation revealed that parameter β_{Change} positively correlated with accuracy score, $r(263) = .38$, $p < .001$ (Figure 8A), but parameter β_{Same} did not, $r(263) = .09$, $p = .146$ (Figure 8B). Furthermore, a Shapiro-Wilk test showed that both variable $\Delta\beta$, $W(263) = .96$, $p < .001$, and the mean of the dual inverse temperatures, $W(263) = .96$, $p < .001$, were non-normally distributed. A Spearman correlation then demonstrated that the accuracy scores correlated positively with both $\Delta\beta$, $r(263) = .32$, $p < .001$ (Figure 8C), and the mean of the dual inverse temperatures, $r(263) = .35$, $p < .001$ (Figure 8D).

In general, these latter results may indicate that the participants who exploit more after an altered state are the participants who are more accurate in this task. However, as we noted earlier, the stimulus stickiness parameter probably interfered with inverse temperature β_{Same} and, thus, it is possible that these correlations were affected by the inclusion of the stimulus stickiness. Therefore, we also decided to perform a *post-hoc* analysis, exploring the correlation between accuracy and the dual inverse temperatures of a model without the stimulus stickiness parameter (i.e., a model with only the dual inverse temperatures and response stickiness as free parameters).

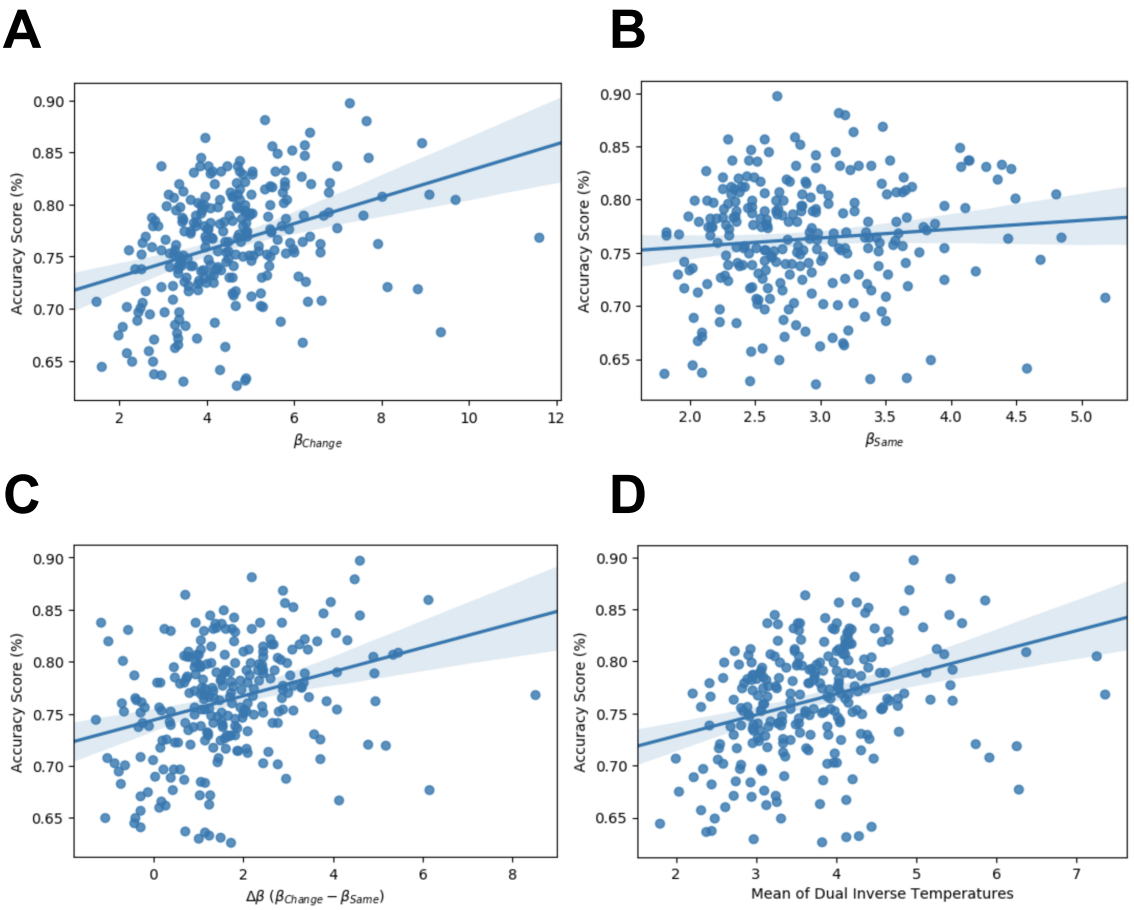

Figure 8. Scatterplots showing the relationships of accuracy with: (A) inverse temperature when the state changed, (B) inverse temperature when the state stayed the same, (C) the difference between the dual inverse temperatures, and (D) the mean of the two inverse temperatures. Here, we used the parameters of the model with the best AIC estimate; a model with dual inverse temperatures, response stickiness, and stimulus stickiness as free parameters. The shading represents the 95% confidence interval.

For the model without stimulus stickiness, we followed the same procedure for the correlation analysis as used for the model with the stimulus stickiness. A Spearman correlation revealed that accuracy correlated positively with both β_{Change}, $r(263) = .38$, $p < .001$ (Figure 9A), and β_{Same}, $r(263) = .29$, $p < .001$ (Figure 9B). For this model, a Shapiro-Wilk test demonstrated that both variable $\Delta\beta$, $W(263) = .96$, $p < .001$, and the mean of the two inverse temperatures, $W(263) = .96$, $p < .001$, were also non-normally distributed. A

Spearman correlation subsequently showed that accuracy did not correlate with variable $\Delta\beta$, $r(263) = .04$, $p = .547$ (Figure 9C), and that there was a significant positive correlation between accuracy and the mean of the inverse temperatures, $r(263) = .37$, $p < .001$ (Figure 9D). The latter results may suggest that participants who have a generally higher exploitation throughout the task are better performers.

Since the best fit model according to AIC had a stimulus stickiness included, the interference of the stimulus stickiness with the inverse temperature of the same states, β_{Same}, may have affected the correlations between β_{Same} and accuracy, and the correlations between $\Delta\beta$ and accuracy. Therefore, we specifically looked at the results of the general inverse temperature β of the best fit model based on BIC, and the results of the dual inverse temperatures of the model without stimulus stickiness. Here, we see that each inverse temperature, as well as the mean inverse temperatures, had a positive correlation with accuracy, except the difference between the dual inverse temperatures, $\Delta\beta$, (of the model without the stickiness parameter) did not correlate with accuracy. Therefore, we can conclude that participants with a generally higher tendency to exploit throughout the task are more optimal decision-makers. These results are in line with our predictions and make sense, as exploration can lead to suboptimal decisions more often than exploitation in our two-step task (due the full reliability of choice rewards and relatively low number of reward changes). So, one way to interpret this is that participants that exploit more during the task are better at the task. However, another way to look at it is that people who (for some reason) are better at the task can afford to exploit more. In this second interpretation, there is no causal relationship between the inverse temperature and accuracy. Note that our dataset does not allow testing of a causal relationship.

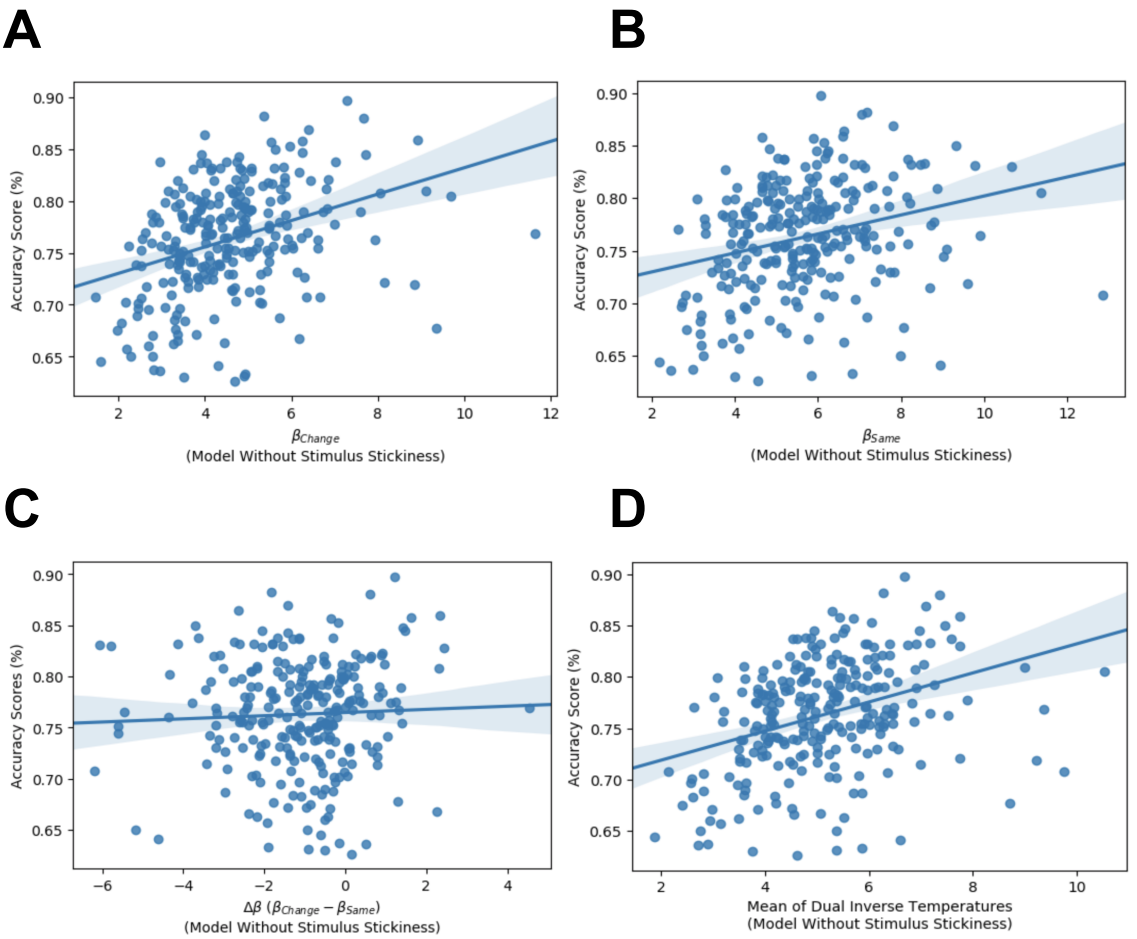

Figure 9. Scatterplots showing the relationships of accuracy with: (A) inverse temperature when the state changed, (B) inverse temperature when the state stayed the same, (C) the difference between the dual inverse temperatures, and (D) the mean of the two inverse temperatures. Here, we used the parameters of the model without stimulus stickiness; a model with only dual inverse temperatures and response stickiness as free parameters. The shading represents the 95% confidence interval.

Discussion

The distinction between goal-directed and habitual behavior has been essential to better understand how humans and other cognitive agents can behave. The reinforcement learning framework has been used to distinguish between goal-directed and habitual behavior, and is often referred to as the distinction between model-based and model-free learning, respectively. The two-step decision-making task is an extensively used paradigm to dissociate between the effects of model-based and model-free learning on behavior, assuming humans use a mixture of these two learning systems (Daw et al., 2011). However, a recent study has shown that people mainly use model-based strategies when the task instructions are very clear in explaining the characteristics of the two-step task (da Silva & Hare, 2020). Moreover, previous studies have shown that certain behaviors in the two-step task are misclassified as model-based or model-free strategies (Akam et al., 2015; da Silva & Hare, 2020). The current study used a model-comparison approach to get a better understanding of how people behave in the two-step task. In general, our results suggest that participants in the two-step task mainly use model-based strategies. In addition, our findings indicate that some of the behavioral tendencies in the two-step task might be better understood at a different dimension of learning or decision-making, such as the trade-off between exploration and exploitation.

Our first research question was whether the participants used a mixture of model-based and model-free strategies in our two-step task, or whether they were pure model-based or pure model-free learners. Our results demonstrated that the participants in our two-step task mainly used model-based strategies. These results are inconsistent with the main assumption that people use a mixture of these two systems during the two-step task (Daw et al., 2011) and are in line with the findings by da Silva and Hare (2020). It is important to note here that, similar to the study by da Silva and Hare (2020), we used task instructions that explicitly explained the structure of the task. These explicit instructions might have helped people understand the model of the task from

the beginning, allowing them to use model-based strategies throughout the task.

In addition, Kool et al. (2016) demonstrated that the modified two-step task also makes people rely more on model-based strategies. Although the results by Kool et al. (2016) showed participants still used a mixture of the two systems, the modified task increasing the reliance on model-based control could also have been an additional reason why we found that participants were mainly model-based learners. Kool et al. (2016) suggested that accuracy-demand trade-off might have motivated participants to rely more on model-based control in the modified task, as this task allows model-based learner to be more accurate, while the original two-step task does not.

However, another possible explanation could be that the adjustments actually simplified the two-step task and, therefore, the task is easier to understand, allowing people to use more model-based strategies. Moreover, da Silva and Hare (2020), Kool et al. (2016), and our study also introduced a story with matching stimuli in the two-step task that was easy to understand (e.g., following an animal in the forest to find a treasure chest), unlike the original task of Daw et al. (2011) where it is very abstract (e.g., Tibetan characters). The latter may also have helped people better understand the task, leading to a higher use of model-based strategies.

Furthermore, previous research has shown that participants use more model-free strategies under a working memory load (Otto, Gershman, et al., 2013). By introducing an easy-to-understand story with matching stimuli, it may have lowered the working memory load. Also, unlike the original task, each choice deterministically led to a specific number of points and there were no choice options in the second-step phase in our task. Our task also had a wider range of reward values and larger reward changes than the original task, making it easier to detect the reward changes and the reward differences between choices. All of these adjustments simplify the task and, thus, may have reduced the working memory load, allowing participants to use more model-based strategies.

Another possible factor of more model-based control in our task is that all our participants were young adults, and previous research showed that young adults use more model-based strategies than older adults (e.g., Ito, Cao, Reinberg, Keller, Monterosso, Schweighofer, & Liew, 2021). Our study did not allow us to investigate to what extent the accuracy-demand trade-off, understanding the task structure, task simplicity, and age of subjects actually affected the reliance on model-based control. Future research is needed to further investigate when model-free control is actually used or not in the two-step task.

Our second research question was whether a different dimension of learning and decision-making may reveal certain behavior in the two-step task, instead of the dimension between model-based and model-free learning. Here, we specifically focussed on the trade-off between exploration and exploitation by using the inverse temperature parameter. We investigated if the tendency to explore changed when the stimuli in the environment (i.e., state) changed. Our findings suggest that people explore the other choice options more often after the stimuli in the environment changed than when the stimuli stayed the same, even though this had no reward-based benefits whatsoever. In the next paragraphs, we will elaborate on some possible interpretations of this higher tendency to explore after a state changed. However, it is important to note here that these findings were part of a *post-hoc* analysis, exploring a computational model without the stimulus stickiness parameter, as the stimulus stickiness can interfere with the exploration-exploitation trade-off.

There are three possible interpretations for why the participants explored more often when the stimuli (state) changed. Before we describe these interpretations, we first want to note that there are two types of exploration: *directed* and *random exploration* (Wilson, Geana, White, Ludvig, & Cohen, 2014). Directed exploration refers to exploration used as a strategy to find an optimal decision, while random exploration refers to accidental exploration due to noise. Thus, directed exploration is reward-based, while random exploration is not. The first possible interpretation is that participants had the false assumption that rewards changed more often when the stimuli changed and

that this false assumption caused them to explore more often to see if the other choice's reward got better or not. In other words, this would be directed exploration based on an incorrect model of the task. Note that this interpretation is in line with the findings by da Silva and Hare (2020) showing model-based strategies can deviate from what is expected from a model-based learner by using an incorrect model of the environment.

A second possible interpretation is that the trials are more demanding when the stimuli change and, therefore, participants selected more often random choices in these trials (i.e., random exploration). During a trial with different stimuli than in the previous trial, the participants have to transition the expected reward from one stimulus to another. This reward transition can be more cognitively demanding because it has to be retrieved from working memory, similar to cases where the task changes (e.g., Rogers & Monsell, 1995). Kool et al. (2010) found evidence that people try to avoid cognitive effort when making decisions. One way to avoid the cognitive effort of a reward transition is by randomly selecting one of the choices, regardless of its expected rewards. Previous studies have shown that people tend to use more random exploration due to higher cognitive load (Cogliati Dezza, Cleeremans, & Alexander, 2019).

However, the results by Cogliati Dezza et al. (2019) suggest that the effects of cognitive load on exploration are due to a decrease in information integration processes instead of more noise in decision-making itself. Thus, a third possible interpretation is that we found that people used more exploration when stimuli changed due to higher cognitive demand, but that this exploration was guided by expected rewards that integrated less information about the environment. In other words, we may have found exploration, which actually corresponds to model-free learning. Previous research has demonstrated that participants use more model-free strategies under a working memory load (Otto, Gershman, et al., 2013). Thus, it is possible that we captured a higher tendency to explore after changing states that actually represents more model-free learning. Our study did not look further into a possible effect of the changed stimuli on the use of model-free strategies. In addition, further

research is needed to investigate which of the two types of exploration are involved at certain points during the two-step task, as our task did not allow us to distinguish between directed and random exploration.

Importantly, our findings suggest that the stimulus stickiness parameter partially absorbed the exploitation from the inverse temperature parameter. This makes sense as the stimulus stickiness and exploitation both show similar responses (i.e., choosing the same choices as in the previous trial). Moreover, the stimulus stickiness specifically represents the tendency to repeat the same stimulus and this is only possible when the same stimuli are presented. Therefore, it makes sense that the stimulus stickiness only absorbed the exploitation of the trials in which the stimuli were the same as in the previous trial. The only difference between the stimulus stickiness and exploitation is that exploitation represents the tendency to repeat the same choices based on their learned reward values, while the stimulus stickiness simply tends to repeat the same choice regardless of the expected rewards. So, it is important here to note that the stimulus stickiness parameter and the inverse temperature parameter interfere with each other, although one is based on the expected rewards while the other is not. Future research should be careful in interpreting these parameters when they are both included in the same model.

Our third research question was whether participants learned the reward values immediately from the outcomes they obtained in our task, which is more optimal in our task. Our results found that the learning rate parameter was better set to its maximum value, which indicates the participants did learn the reward values quickly. The reward outcome of a stimulus was completely reliable in our task because choosing a stimulus deterministically led to a specific reward. When a learning rate is set to its maximum value, an agent completely changes the expected reward of the stimulus into the previous perceived outcome of the stimulus. Therefore, changing the expected rewards of stimuli completely into the values of the last obtained outcomes is more optimal in our two-step task, as well as easier to do compared to the original two-step task that used reward probabilities.

Our fourth question was whether individuals who relied more on model-based strategies were more accurate. We used the modified two-step task because it should maximize the benefits of model-based control. Additionally, Kool et al. (2016) found that a higher use of model-based control indeed correlated with accuracy in their modified two-step task, on which we based ours. In contrast with Kool et al. (2016), our study did not find a relationship between model-based control and accuracy scores. However, the participants in our study mainly used model-based control. The high values of the balance parameter between the model-based and model-free systems clearly reached a ceiling effect, and this may be the reason why we did not find a correlation between the balance parameter and accuracy scores.

As a fifth research question, we were interested if accuracy was related to other behavioral responses in our task, namingly the trade-off between exploration and exploitation. Here, we found that the participants that generally exploited more had higher accuracy scores. This can be interpreted in two ways. The first interpretation is the causal explanation that exploration more often leads to suboptimal decisions than exploitation in our two-step task. This makes sense because the choice rewards were completely reliable in our task and there were relatively low numbers of reward changes throughout the task. These two latter factors make explorative decisions more likely to be suboptimal.

The second interpretation is a non-causal relationship that people who make more optimal decisions can afford to exploit more. For example, it is possible that some participants were lucky by often getting high reward values from their selected choices and that this allowed them to continue exploiting these rewards without having to explore to find better rewards. For instance, when an agent receives a reward of 9 points when choosing a stimulus, (s)he does not have to explore the other choice option because her/his current choice option is most likely to be the most optimal choice (when the upper bound of the reward values is 10 points). This was not controlled for in the current study, because the possible number of points that could be earned were randomly assigned to the participants. Here, we would like to address that the trade-off

between exploration and exploitation is one of the major dilemmas (i.e. seeking information or maximizing rewards) in behavioral science (Cohen et al., 2007), especially important in understanding how and when people change behavior to make more optimal decisions and what the individual behavioral differences are. However, this trade-off still remains unclear. Future research should further explore how this dilemma is involved in learning and decision-making.

To conclude, the current study suggests that people use mainly model-based strategies when performing the two-step task, which probably depends on how well the participants understand the task structure. On the other hand, our results suggest that the exploration-exploitation trade-off plays an important role in how people behave in the two-step task, and could explain some behavioral tendencies more than what we are able to explain by simply looking at the distinction between the model-based and model-free system. A better understanding of the behavior in the two-step task, beyond the distinction between model-based and model-free learning, should be addressed by future research. Furthermore, future research should carefully consider several factors when using the two-step task, such as how explicit the task instructions are, how easy the task is, and whether the task structure avoids misclassification of model-based and model-free learning.

References

Akam, T., Costa, R., & Dayan, P. (2015). Simple Plans or Sophisticated Habits? State, Transition and Learning Interactions in the Two-Step Task. *PLoS Computational Biology, 11*(12), e1004648. https://doi.org/10.1371/journal.pcbi.1004648

Cogliati Dezza, I., Cleeremans, A., & Alexander, W. (2019). Should we control? The interplay between cognitive control and information integration in the resolution of the exploration-exploitation dilemma. *Journal of Experimental Psychology: General, 148*(6), 977–993. https://doi.org/10.1037/xge0000546

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 of London. Series B, Biological sciences, 362*(1481), 933–942. https://doi.org/10.1098/rstb.2007.2098

Collins, A. G., & Cockburn, J. (2020). Beyond dichotomies in reinforcement learning. *Nature Reviews Neuroscience, 21*(10), 576-586. https://doi.org/10.1038/s41583-020-0355-6

da Silva, C. F., & Hare, T. A. (2020). Humans primarily use model-based inference in the two-stage task. *Nature Human Behaviour, 4*(10), 1053-1066. https://doi.org/10.1038/s41562-020-0905-y

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69*(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience, 8*(12), 1704–1711. https://doi.org/10.1038/nn1560

Daw, N. D. (2018). Are we of two minds? *Nature neuroscience, 21*(11), 1497-1499. https://doi.org/10.1038/s41593-018-0258-2

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The Good, The Bad and The Ugly. *Current Opinion in Neurobiology*, *18*(2), 185–196. https://doi.org/10.1016/j.conb.2008.08.003

Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., & Daw, N. D. (2015). Model-based choices involve prospective neural activity. *Nature Neuroscience, 18*(5), 767–772. https://doi.org/10.1038/nn.3981

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: carving nature at its joints. *Current opinion in neurobiology*, *20*(2), 251–256. https://doi.org/10.1016/j.conb.2010.02.008

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology, 71*, 1–6. https://doi.org/10.1016/j.jmp.2016.01.006

Heitz, R. P. (2014). The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in Neuroscience, 8*, 150. https://doi.org/10.3389/fnins.2014.00150

Ito, K. L., Cao, L., Reinberg, R., Keller, B., Monterosso, J., Schweighofer, N., & Liew, S. L. (2021). Validating Habitual and Goal-Directed Decision-Making Performance Online in Healthy Older Adults. *Frontiers in aging neuroscience, 13*, 702810. https://doi.org/10.3389/fnagi.2021.702810

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Computational Biology, 7*(5). https://doi.org/10.1371/journal.pcbi.1002055

Kool, W., Cushman, F. A., & Gershman, S. J. (2016). When Does Model-Based Control Pay Off? *PLoS Computational Biology, 12*(8), 1–34. https://doi.org/10.1371/journal.pcbi.1005090

Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General, 139*(4), 665–682. https://doi.org/10.1037/a0020198

Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural Computations Underlying Arbitration between Model-Based and Model-free Learning. *Neuron, 81*(3), 687–699. https://doi.org/10.1016/j.neuron.2013.11.028

Otto, A. R., Gershman, S. J., Markman, A. B., & Daw, N. D. (2013). The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological Science, 24*(5), 751–761. https://doi.org/10.1177/0956797612463080

Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A., & Daw, N. D. (2013). Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences of the United States of America, 110*(52), 20941–20946. https://doi.org/10.1073/pnas.1312011110

Piray, P., Toni, I., & Cools, R. (2016). Human Choice Strategy Varies with Anatomical Projections from Ventromedial Prefrontal Cortex to Medial Striatum. *Journal of Neuroscience, 36*(10), 2857–2867. https://doi.org/10.1523/JNEUROSCI.2033-15.2016

Radenbach, C., Reiter, A. M., Engert, V., Sjoerds, Z., Villringer, A., Heinze, H. J., Deserno, L., & Schlagenhauf, F. (2015). The interaction of acute and chronic stress impairs model-based behavioral control. *Psychoneuroendocrinology, 53*, 268–280. https://doi.org/10.1016/j.psyneuen.2014.12.017

Rescorla, R. A., & Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A.H. Black & W.F. Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory* (pp. 64–99). New York: Appleton-Century-Crofts. https://pdfs.semanticscholar.org/afaf/65883ff75cc19926f61f181a687927789ad1.pdf

Rogers, R. D., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General, 124*(2), 207–231. https://doi.org/10.1037/0096-3445.124.2.207

Shapiro, S. S., & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika, 52*, 591-611. https://doi.org/10.1093/biomet/52.3-4.591

Skinner, B. F. (1938). The behavior of organisms: An experimental analysis. In R.M. Elliot (Ed.), *The Century Psychology Series* (pp. 1–451). New York:

Appleton-Century. https://www.scribd.com/document/283214535/Skinner-B-F-1938-the-Behavior-of-Organisms-An-Experimental-Analysis

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction.* MIT press. http://incompleteideas.net/book/RLbook2020.pdf

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements, 2*(4), i–109. https://doi.org/10.1037/h0092987

Tolman, E.C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208. https://doi.org/10.1037/h0061626

Verguts, T., & Storms, G. (2004). Assessing the informational value of parameter estimates in cognitive models. *Behavior Research Methods, Instruments, & Computers, 36*(1), 1-10. https://doi.org/10.3758/BF03195544

Wilcoxon, F. (1945) Individual Comparisons by Ranking Methods. *Biometrics Bulletin, 1*, 80-83. https://doi.org/10.2307/3001968

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General, 143*(6), 2074–2081. https://doi.org/10.1037/a0038199