The dual-systems approach in reinforcement learning – a multimethodological analysis of strategies, mechanisms, and modulating factors

Dissertation

zur Erlangung des Doktorgrades

an der Universität Hamburg,

Fakultät für Psychologie und Bewegungswissenschaft,

Institut für Psychologie

vorgelegt von

Anna Cremer

Hamburg, 2024

Tag der mündlichen Prüfung: 07.10.2024

Promotionsprüfungsausschuss

Vorsitzender: Prof. Dr. Martin Spieß

1. Dissertationsgutachter: Prof. Dr. Lars Schwabe
2. Dissertationsgutachter: Dr. Jan Gläscher

1. Disputationsgutachter: Prof. Dr. Ulf Liszkowski
2. Disputationsgutachterin: Prof. Dr. Anja Riesel

Acknowledgements

I would like to express my gratitude to Prof. Dr. Lars Schwabe for his guidance, invaluable support, and endless patience throughout the entirety of my doctoral journey. His expertise and mentorship have been instrumental in shaping both the direction and quality of this research. I am also profoundly thankful to Dr. Jan Gläscher for introducing me to the fascinating world of computational modeling. Our collaboration in my first PhD project set the course for my subsequent research.

My heartfelt appreciation extends to the entire lab team for their camaraderie, support, and countless joyful moments we shared. I thank the first generation of PhD candidates with Lisa, Lisa, Lisa, and Mario for being great role models and big-hearted supporters. My deepest gratitude further goes to Conny – your expertise, your friendship, and your energetic nature meant a lot to me. Nadine, Franzi, Gundi, it was a pleasure to be you colleague and friend and I could not have asked for a greater company through this journey. Felix and Meike, thank you for being such a great extension to our group. Stefan, Valentina, and Hendrik, thank you for being a great team. Anna-Maria, I especially thank you for supporting me in the last steps to complete the submission of this thesis. Words can't express the gratitude I feel for you, Felix, for being the best office partner, mentor, supporter, and friend I could wish for. Anke, your enthusiastic, dynamic and humorous nature has been invaluable for me.

Lastly am indebted to my friends and family for their unwavering support, understanding, and for always being there during the highs and lows of this academic journey. There are too many to name, but I am deeply grateful to having all of you in my life. Your friendship and encouragement have been a source of strength and motivation.

This dissertation would not have been possible without the collective support, encouragement, and inspiration of all those mentioned above. Thank you.

I dedicate this thesis to my dear friend Felix Kruse.

Abstract

Reinforcement learning offers a formalized model of decision-making processes guided by the predicted values associated with available options. It is commonly assumed in these models that behavior is governed by two separate systems: a fast, reflexive system and a slower, more deliberate prospective system. These systems are thought to be dissociable not only on the behavioral, but also on the neural level. At the same time, dual-systems approaches have been criticized for their allegedly oversimplistic nature. It is argued that cognitive processes are better reflected on a continuum and that the brain uses more integrated, dynamic, and context-dependent mechanisms than captured in dichotomized systems. To address these issues, we tested the notion of dual-systems approaches underlying RL, probing the presence of separate systems, their interplay, and modulating factors. Stress is known to be a powerful modulator of leaning and decision making that was shown to induce a shift from cognitively demanding to rather reflexive systems. We examined three different dual-systems frameworks in RL and tested the susceptibility of the systems and their components to stress and stress mediators. In Study 1, we examined the extent to which adaptive behavior is driven by a purely reward-driven model-free reinforcement learning, versus a model-based strategy that incorporates a map of the environment to guide choices, while exposing participants to an acute stress manipulation. Although stress is assumed to impair prefrontal functions associated with model-based reinforcement learning, participants from the stress and control groups utilized both learning strategies in a stimulus-response association task with an overall bias towards model-free reinforcement learning. However, our results from functional magnetic resonance imaging showed a reduction of value computations underlying both model-free and model-based reinforcement learning in stressed participants. In Study 2, we aimed to shed light on the processing of specific choice components underlying the preference for exploiting known, but depleting resources versus exploring unknown options. Prior research has identified dopamine and noradrenaline as key drivers in this tradeoff. By pharmacologically blocking either of these two neurotransmitter systems, we found that neither of them drives exploration vs. exploitation per se. Rather, they both play functionally different roles: While dopamine signaled choice-relevant features. noradrenaline exerted a disengagement from the current information path. In Study 3, we tested how the systems of working memory vs. reward learning contribute to the acquisition of stimulus-action pairs and whether these contributions are subject to stress effects. Our results show a cooperative interplay between working memory and reinforcement learning with reward learning guiding behavior when working memory limits are exceeded. Overall, our findings challenge the strict dichotomy traditionally posited in dual-systems theories, highlighting a more nuanced interplay of cognitive components in learning and decision-making.

Therefore, this work contributes to a deeper understanding of the adaptive nature of human cognition and offers implications for enhancing decision-making strategies in real-world scenarios.

Contents

List of Figures

Figure 2, and the parts depicting neural mechanisms in Figures 4,7, and 9 were created with BioRender.com.

List of Abbreviations

dACC	dorsal anterior cingulate cortex
dlPFC	dorsolateral prefrontal cortex
EEG	electroencephalography
fMRI	functional magnetic resonance imaging
IPS	intraparietal sulcus
MFG	medial frontal gyrus
MVT	marginal value theorem
OFC	orbitofrontal cortex
PFC	prefrontal cortex
PE	prediction error
RL	reinforcement learning
RLWM	reinforcement learning working memory
RPE	reward prediction error
SFG	superior frontal gyrus
SPE	state prediction error
TD	temporal difference
vmPFC	ventromedial prefrontal cortex
WM	working memory

1 Introduction

Learning from reward is one of the key principles underlying decision-making and behavioral adaptation. It is based on the fundamental observation that actions that have been rewarded are more likely to be repeated than ones that have not been rewarded, or even punished. This led to the development of the reinforcement learning (RL) approach, in which learning and decision-making behavior is guided by the reward history obtained for available options in the current decision space (Sutton & Barto, 1998). The objective is to learn a policy – a strategy for selecting actions – that maximizes cumulative rewards over time (Gershman & Uchida, 2019). The key feature of RL lies in its ability to translate observed behaviors into mathematical formulations, thereby offering a precise model for quantifying how an individual learns from interactions with the environment (Collins & Cockburn, 2020). Strikingly, these formulas can not only reflect behavior, but can also be applied to the underlying neural mechanisms. RL therefore provides a normative framework to explain how the brain processes information, adapts to changing circumstances, and optimizes behavior to achieve desired outcomes (Dayan & Balleine, 2002).

The empirical study of learning and decision-making has raised a variety of seminal dichotomic approaches based on the idea that behavior is controlled by two dissociable systems (Sloman, 1996; Stanovich & West, 2000; Tversky & Kahneman, 1974). Typically, these dual-systems frameworks contrast a fast, reflexive system that spares resources with one that incorporates more complex information but is slower and computationally heavy. Examples of such dual-systems frameworks in RL include goal directed vs. habitual behavior (Balleine & O'Doherty, 2010), model-free vs. model based RL (Gläscher et al., 2010), or the exploration-exploitation dilemma (Blanchard & Gershman, 2018). These approaches often differ in their specific emphasis on individual aspects but share core computational principles and overlap in the underlying neural mechanisms.

In spite of the assumed dichotomization, the most prominent dual-systems frameworks in RL suggest that the respective systems overlap to some point as well on the behavioral as on the computational and neural levels. This indicates that behaviors are not solely the result of one of two systems but a combination of both (Drummond & Niv, 2020; Wilson et al., 2021). Moreover, the dimensionality of learning extends beyond a mere binary classification, hinting that such frameworks, while useful, may omit valuable insights by simplifying the complexity of learning processes into two categories (Collins & Cockburn, 2020). Therefore, dual-systems theories have been criticized for promoting approximative solutions to complex problems in cognitive science (Gigerenzer, 2010; Kruglanski & Gigerenzer, 2011). In a more global perspective, dual-systems theories might be not suitable to provide an integrative view on cognitive processing, both on a behavioral and on a neural level.

In this thesis, we highlight three dichotomization frameworks that are popular in the field of RL, namely model-free vs. model-based RL, exploration vs. exploitation, and working memory (WM) vs. reward learning. We propose a critical perspective on the dichotomization of RL processes by identifying the interplay of the proposed two systems within these domains. Further, we test whether modulating factors alter individual process components and how this modulation affects the interplay between systems.

1.1 Reinforcement learning

The finding that previously rewarded actions are more likely to be repeated than ones that have not been rewarded or even punished (Thorndike, 1927) was first manifested in the paradigms of Pavlovian and instrumental conditioning and has accumulated a broad range of evidence since (Balleine & Ostlund, 2007; Bouton et al., 2021). A striking observation was that it is not the reward per se that reinforces behavior, but the difference between a predicted value of future rewards and the actual outcome (Schultz et al., 1997). This concept goes back to the Rescorla-Wagner model, a classic computational model originally formulated to explain associative learning in Pavlovian conditioning (Rescorla & Wagner, 1972). It proposed that the change in strength of a stimulus-outcome association is determined by the discrepancy between the expected and actual outcomes. Learning therefore occurs whenever there is an unpredicted event, and the amount of learning is proportional to the surprise associated with the outcome. These prediction errors (PEs) are assumed to be the central driver of incremental learning. The foundational theoretical basis was applied to RL, where learning was driven by the difference between the predicted reward for an option and the actual outcome, known as the reward prediction error (RPE; Schultz et al., 1997). Specifically, a trial-unique RPE δ_t updates the value $V(a)_t$ that is associated with choosing a specific action a, based on the observed reward *R(a):*

$$V(a)_{t+1} = V(a)_t + \alpha\delta_t \tag{1}$$

$$\delta_t = R(a)_t - V(a)_t \tag{2}$$

for trial *t*, where α with $0 < \alpha < 1$ indicates the learning rate. Hence, the RPE signals whether the outcome from choosing a specific action was better or worse than predicted, thereby determining the direction of the behavioral adaptation (Daw & Doya, 2006; Pessiglione et al., 2006; Schönberg et al., 2007). The mismatch between the expected outcome and the actual outcome is then used to improve the prediction, according to a learning rate α which controls the degree to which extent the prediction error leads to an adjustment of action values.

Taking this as a basis, RPEs further serve to iteratively refine policies towards an optimal value function. Temporal difference (TD) approaches offer a formal description of how RPEs enable learning incrementally about associative relationships between actions and outcomes (Sutton & Barto, 1998). The TD-RPE is calculated by first adding the immediate reward obtained in the current trial $R(s)_t$ to the expected reward for this option $V(s)_{t+1}$, which is an integration of all possible future values discounted by a factor γ. From this sum, the previous, not yet updated reward estimate for the chosen option $V(s)_t$ is subtracted to compute the TD-RPE:

$$\delta_t = R(s)_t + \gamma V(s)_{t+1} - V(s)_t \tag{3}$$

At the core of TD learning lies the idea that learning is about identifying a value function that guides behavior towards the most advantageous options to maximize rewards in the long run (Gershman & Uchida, 2019). This value function contains all information necessary to make a choice, such as an environment with a defined set of states $\boldsymbol{S}$, a set of actions $\boldsymbol{A}$ available in these states, and rewards $\boldsymbol{R}$ associated with the actions. The agent finds itself in a state $\boldsymbol{s} \in \boldsymbol{S}$ and it can change its current state by choosing an action $\boldsymbol{a} \in \boldsymbol{A}$, following the environment's state transition structure $\boldsymbol{P}$. Subsequent actions $\boldsymbol{a} \in \boldsymbol{A}$ are probabilistically associated with rewards $\boldsymbol{R}$. The reward and state transition probability distributions specify how state-action pairs lead to rewards and new states, respectively (Figure 1). The agent's goal is to find a policy that maximizes the value **Q** (Sutton & Barto, 1998).

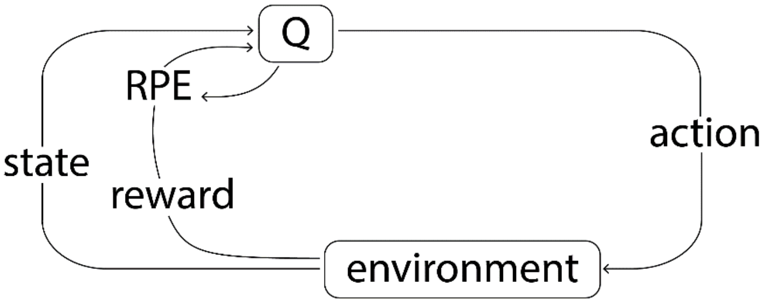


Figure 1. Basic conceptualization of reinforcement learning problems. In an environment with a defined set of states, the agent can choose between a defined set of actions that are probabilistically associated with rewards. A value Q is computed for each choice option as a function of state transition and reward probability. Q is updated after each encounter with the option, based on the RPE. Reprinted with permission from Yoo & Collins (2021).

Within this framework the learning process can be subdivided in three steps: (1) predicting the reward value of currently available actions, (2) selecting the action with the maximum reward value, and (3) updating the predictions based on experience (Daw & Doya, 2006).

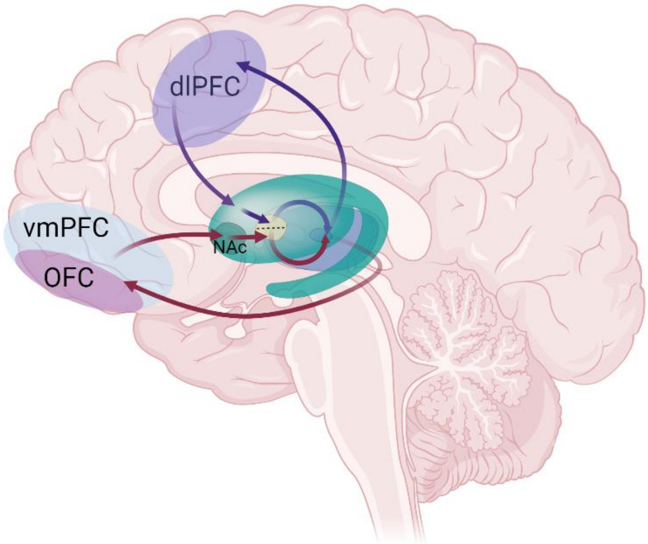


Figure 2. Schematic representation of reinforcement learning in the brain. Ventral and dorsal striatal dopamine support learning about state values and action policies, respectively. The ventral path (red lines) processes information from ventral prefrontal areas (vmPFC and OFC) to the ventral striatum, projecting to the ventral pallidum (yellow, bottom part). In the dorsal path (purple lines), the dorsal striatum receives input from the dlPFC with is projected to the dorsal pallidum (yellow, top part). Both ventral and dorsal pallidum send projections back to the medial-dorsal nucleus of the thalamus which then sends topographically organized return projections to the vmPFC and the dlPFC, closing the loop (Averbeck & O'Doherty, 2021).

Importantly, RPEs are reflected in phasic firing of dopaminergic neurons in the ventral tegmental area (Bayer & Glimcher, 2005; Roesch et al., 2007; Schultz et al., 1997). Upon transmission to the striatum and cortex, the RPE signals are used to update stimulus-action values (Wickens et al., 2007). Moreover, RPEs drive synaptic plasticity in the striatum and thereby translate learned associations into behavioral policies. Choices that led to positive RPEs are reinforced, and those that led to negative RPEs are weakened (Frank et al., 2004; Reynolds & Wickens, 2002; Shen et al., 2008). Specifically, the ventral striatum (comprising the ventral putamen and nucleus accumbens) is associated with dopamine learning of state values (Knutson & Cooper, 2005; Pagnoni et al., 2002; Pessiglione et al., 2006), while the dorsal striatum (comprising the nucleus caudatus and dorsal putamen) supports learning about action values (Daw et al., 2005a; O'Doherty et al., 2004). This dissociation of ventral vs. dorsal striatum is extended to cortical regions (Figure 2) with ventral cortical regions like the ventromedial prefrontal cortex (vmPFC) and the orbitofrontal cortex (OFC) being associated with processing stimulus-outcome associations (Camille et al., 2011; Ostlund & Balleine, 2007; Rudebeck et al., 2008; Rushworth et al., 2012) and goal-related features (O'Doherty, 2011; Padoa-Schioppa & Assad, 2006) while dorsal cortical regions like the dorsolateral prefrontal cortex (dlPFC) were found to be involved in action-related processes (Platt & Glimcher, 1999; Quilodran et al., 2008; Sugrue et al., 2004).

Essentially, RL occurs in interplay with various cognitive components of decision-making. For example, it interacts with *working memory* (Collins & Frank, 2012), which enables the retention of behaviorally relevant information, *motivation* (Dayan & Balleine, 2002), which influences decision-making priorities, and *motor processing* of chosen actions (Barto, 2003). Moreover, it incorporates *episodic memory*, allowing the system to draw from past experiences rich in perceptual details, which enables effective learning particularly in complex real-world-scenarios (Gershman & Daw, 2017).

1.2 Dual-systems approaches

Dual-systems approaches are attractive frameworks for narrowing down the space of computational solutions for decision-making problems to two alternative (and sometimes mutually exclusive) computational mechanisms. In RL in neuroscience and psychology, there have been a range of such dichotomies. The most prominent ones include model-based vs. model-free RL, exploration vs. exploitation strategies, and reward learning vs. cognitive processing, specifically WM. In the following, we will discuss both supporting evidence and criticisms of these dualistic conceptualizations.

1.2.1 Model-based vs. model-free RL

The model-based vs. model-free dichotomy is one of the most popular dual-systems distinctions in RL, originating the idea that it is guided by two different strategies that rely on dissociable neural mechanisms. *Model-based RL* relies on an explicit model of the environment that allows planning and simulating actions, while *model-free RL* learns from direct experience without building an internal model (Daw et al., 2011; Dolan & Dayan, 2013; Gläscher et al., 2010). Take as an example that you are on your way home from work. While sitting in the train, an announcement proclaims that the ride will be delayed for an unknown period of time. A model-based strategy would make use of its cognitive map of the route network, coming to the conclusion that taking an alternative route will take nearly the same time as the usual route, despite changing trains at one point. The model-free agent on the other hand, acting based on experience, would prefer alternatives routes that were successful in the past.

Behavioral Systems

Both strategies use prediction errors to learn the most advantageous option, but they vary in the strategy or "policy" underlying action selection. The model-free strategy acts retrospectively – it learns values by trial and error and stores them as a set of value estimates, each representing the integrated reward history for options chosen in the past, as described in 1.1. A model-based agent on the other hand forms a cognitive map of environmental contingencies to evaluate future possibilities in a prospective manner. It considers not only the outcome of an action, but also state-transition and reward functions, that is, representations of all state-action sets and the probabilistic dependencies between their elements (Daw et al., 2005b, 2011; Gläscher et al., 2010).

To dissociate the two learning strategies in an experimental setting, a sequential Markov decision task, also known as a two-step task, was developed (Daw et al., 2011; Gläscher et al., 2010). The task consists of two successive stages, followed by an outcome (Figure 3A, B). In the first stage (state 1) the agent can choose between two options (see Figure 3A, yellow state). Each option is predominantly connected to one out of two second stage states – state 2 (blue) or state 3 (red). From there, agents again have the choice between two options, each probabilistically associated with a reward. The goal is to find a policy that maps each state to the action with maximum expected reward. Model-based and model-free RL strategies predict different first stage choice patterns as a function of previous-trial reward. In model-free learning, choices are driven solely by the reward history without taking the transition paths into account. The model-based learner, on the other hand, would build a model of the task's state transition structure and choose according to the highest joint probability of receiving a reward (Figure 3C). Model-based learning therefore comprises two processes: learning the task structure within states and the transitions between states (state learning), followed by the learning of the value of the second stage states (state value learning; Doody et al., 2022). This is also reflected in a different formalization of the learning signal guiding behavioral adaptation. While model-free RL is solely driven by the reward and therefore learns through RPEs, model-based learning links reward information to estimates of the transition probabilities. The system uses a state prediction error (SPE) to update the cognitive map – in particular, to acquire *state-action-state-transition* probabilities (Gläscher et al., 2010). Empirical data shows that humans' choice behavior in the two-step task is best captured by a mixture of model-based and model-free learning strategies, rather than pure model-free or pure model-based learning (Figure 3C; e.g., Daw et al., 2011; Deserno et al., 2015; Doll et al., 2016; Doody et al., 2022; Gläscher et al., 2010; Otto et al., 2013; Smittenaar et al., 2013; Wunderlich et al., 2012).

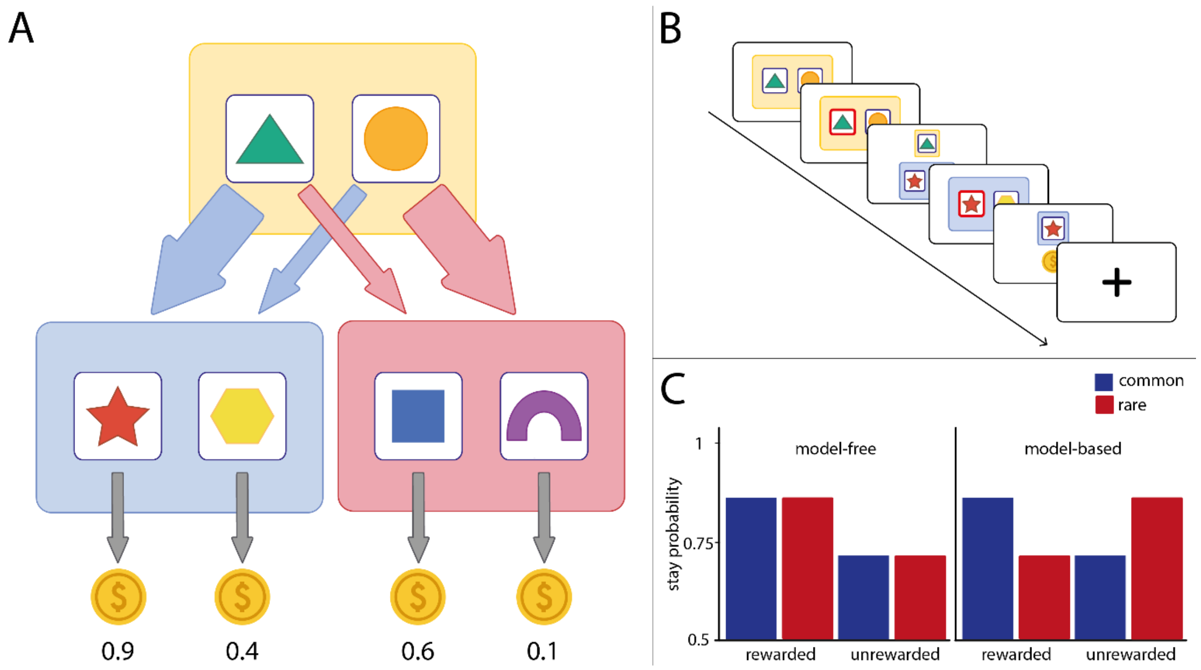


Figure 3. Two step task to dissociate model-based ad model-free reinforcement learning. A) State transition structure. Each first stage (yellow) action is predominantly associated with one of two second stage states (blue or red). The choices in the second stage probabilistically lead to a reward. B) Example trial sequence. The agent chooses the left stimulus that commonly leads to the blue state. At the second stage, the stimulus with the highest reward probability is chosen, followed by a reward. C) Factorial analysis of choice behavior. A model-free strategy predicts choosing an option that was previously rewarded, regardless of whether that reward occurred after a common or rare transition. Choices made by a model-based learner result from an interaction of reward and transition probability. Adapted with permission from Cremer et al. (2021).

Computational Systems

The assumptions of the model-based and model-free learning systems were formalized in a computational model (Figure 4, middle, Daw et al., 2011; Gläscher et al., 2010), where choices derive from a weighted combination of both model-free and model-based value computations. At the core of this model is the concept of a value function associated with each stimulus-action pair across the two stages of the task. For model-free learning, the model updates the value of state-action pairs retroactively based on whether those actions were rewarded in preceding trials, employing the RPE as fundamental mechanism. For model-based learning, the value function is first updated based on the SPE mirroring the discrepancy between expected and actual transitions for the first stage state-action pairs and then combined with the second stage's reward predictions, i.e. the RPE. A critical feature of the model is the weighting parameter w, which quantifies the relative influence of model-based and model-free value computations on behavior. Learning rates α_1 and α_2, estimated for both stages, control the degree of how new information updates the value of actions for both model-based and model-free strategies. Additionally, inverse temperature parameters β_1 and β_2, ranging from 0 to ∞, reflect the extent to which decisions are influenced by learned values in both stages, with lower values indicating rather random of choices and higher values reflecting rather deterministic decisions.

Neural Systems

The dual-systems identified in the two-step task and the hybrid learning model were further shown to map on separate neural systems. In line with the previously described dopaminergic RPE firing (see 1.1) the key region associated with model-free RL is the ventral striatum (Gläscher et al., 2010; Glimcher, 2011; Pagnoni et al., 2002). Further, correlates of reward prediction were repeatedly found in the vmPFC (Averbeck & O'Doherty, 2022; Jocham et al., 2011; O'Doherty, 2011), pointing towards an interaction of dopaminergic RPE signaling in the ventral striatum and the vmPFC using these signals to anticipate upcoming rewards in model-free RL. The dopamine system is also involved in model-based RL (Daw et al., 2011; Deserno et al., 2015; Sadacca et al., 2016; Sharp et al., 2016; Wunderlich et al., 2012). Evidence from a genetic study suggests that the *striatal* dopamine system is closely linked to model-free RL, while *prefrontal* dopamine was associated with model-based RL (Doll et al., 2016). Moreover, model-based SPEs were further found in the intraparietal sulcus (IPS) and the dlPFC (Daw et al., 2011; Gläscher et al., 2010; Lee et al., 2014; Möhring & Gläscher, 2023). Subsequent studies highlighted the role of the hippocampus as a predictive map mirroring goals and relations between states (Garvert et al., 2017; Miller et al., 2017; Pfeiffer & Foster, 2013; Stachenfeld et al., 2018). Further, the inferolateral prefrontal cortex (ilPFC) has been associated with the role of an arbitrator, signaling whether behavioral control should be controlled by the model-based or the model-free RL system (Lee et al., 2014). In summary, model-free RL is primarily driven by a ventral pathway with the key regions vmPFC and ventral striatum (Figure 4, right), while key regions implicated in model-based RL are the dlPFC, the dorsal striatum, and the IPS (Figure 4, right).

Conclusion

Together, a large body of work suggest that model-based and model-free learning strategies rely on neurally and cognitively dissociable systems. Yet, these systems are likely to interact and share common mechanisms. For example, just as in model-free RL, model-based RL incorporates representations of stimulus-action-reward associations. The degree to which the two systems differ, or the extent to which they arise from the same processes, is largely unclear. In particular, a separation of the systems at the neural level does not seem as sharp as the concept of dual mechanisms suggests.

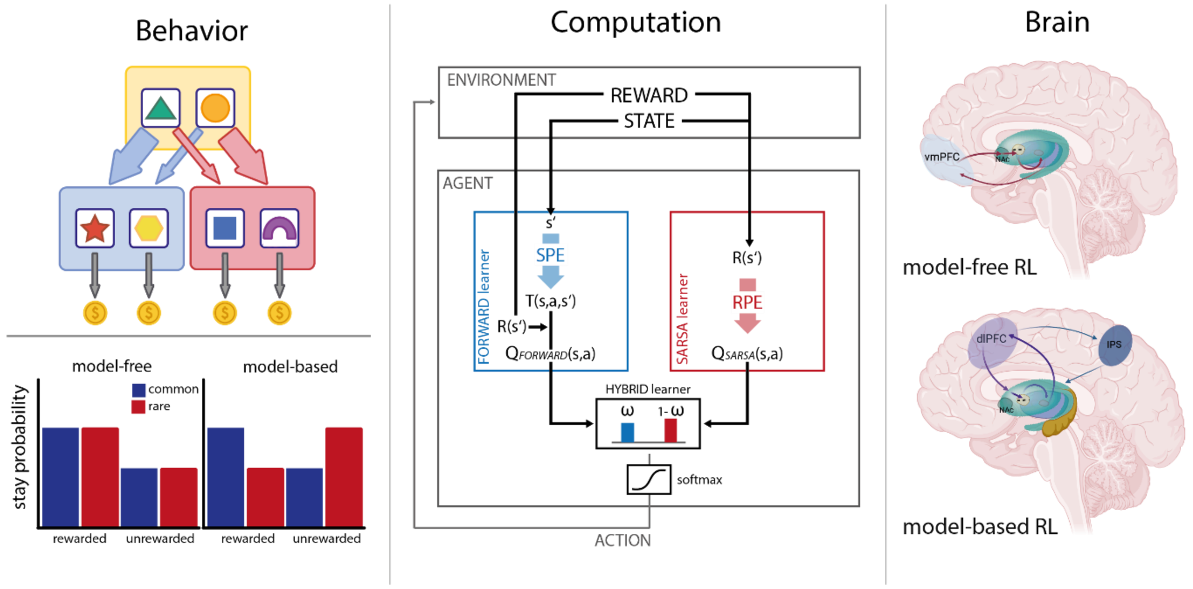


Figure 4. Summary of the model-based vs. model-free dichotomy. Left: Model-based and model-free RL predict different patterns of choice behavior in the two-step task (top). Model-free RL predicts that a previously rewarded action is more likely to be repeated on the subsequent trial, independent of the transition. Model-based behavior considers both rewards and transitions: a reward obtained via a rare transition predicts a switch to the other first stage option (bottom). Middle: Computational modelling offers a formal approach, assuming that decisions are driven by value functions that are differentially computed by both systems. Model-based learning uses a combination of the SPE and the RPE to update value functions, while model-free learning performs value updates based on the RPE (reprinted from Möhring & Gläscher, 2023 with permission). Right: Key regions associated with model-free RL were the ventral striatum and the vmPFC (top), while signature regions of model-based RL were the dlPFC, IPS, and hippocampus. Like model-free RL, model-based RL also uses striatal RPEs to guide behavior (bottom).

1.2.2 Exploration vs. exploitation

Another decision-making problem that occurs within the RL framework is the question of when to abandon a known option in favor of a potentially better, but unknown alternative. In classic RL models, a "fully greedy" agent chooses according to a regularly updated record of expected values for each option (Sutton & Barto, 1998), thus exploiting known, reliable options to maximize short-term rewards. In real-world situations, this approach raises two problems. First, it is not possible to know the values of all available options without exploring them at some point. To collect this information, the agent has to occasionally explore new options at the expense of temporarily choosing less rewarding alternatives (Schulz & Gershman, 2019). Second, values are not necessarily stable over time – for example, a job could have been perfect five years ago, but with increasing experience there might be a better fit. The exploration system is associated with seeking novel options and gathering information in order to maximize long-term rewards. But how does the brain know when to explore? Living in a complex and dynamic world, flexibly adapting to new circumstances requires a careful balance between the exploitation of known resources for reward, and the exploration of new opportunities for information (Zajkowski et al., 2017). Extensive exploitation fosters inflexibility, while over-exploration may lead to inefficient and inconsistent behavior, thereby reducing long-term payoffs (Cohen et al., 2007; Gershman, 2018; Mehlhorn et al., 2015). How humans resolve this dilemma is largely unknown, despite its relevance for everyday life and in the context of mental health. An extreme avoidance of uncertainty for example is one of the main characteristics of obsessive-compulsive disorders (Marzuki et al., 2021), whereas impulsive and unstable behavior is associated with attention-deficit/hyperactivity disorder (Hauser et al., 2014). Therefore, a better understanding of the exploration-exploitation-tradeoff is of past and ongoing interest.

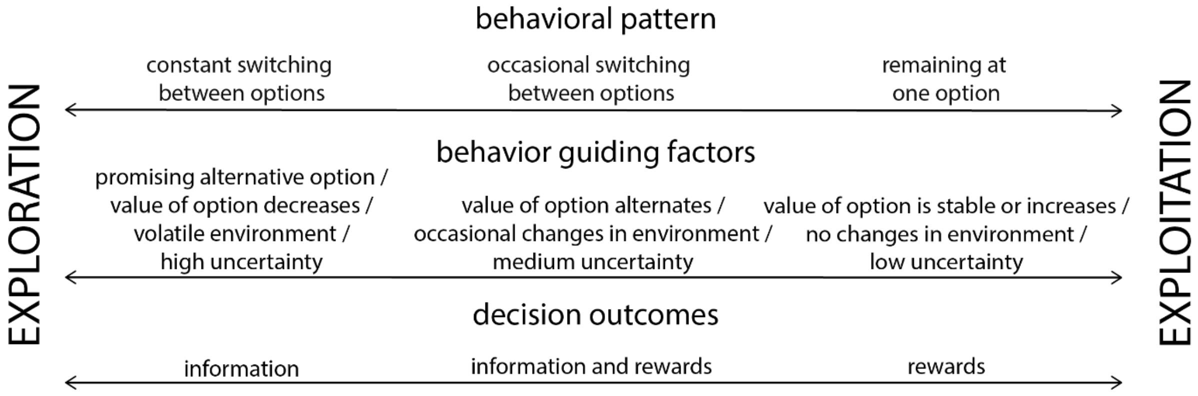

EXPLORATION
behavioral pattern
constant switching between options
occasional switching between options
remaining at one option
behavior guiding factors
promising alternative option / value of option decreases / volatile environment / high uncertainty
value of option alternates / occasional changes in environment / medium uncertainty
value of option is stable or increases / no changes in environment / low uncertainty
decision outcomes
information
information and rewards
rewards
EXPLOITATION

Figure 5. **Behavioral signatures of exploration and exploitation.** Top: Switching to new options vs. staying at one option are the behavioral patterns reflecting exploration vs. exploitation. Middle: Value changes, stability of the environment, and uncertainty are the most important factors driving explorative vs. exploitative decisions. Bottom: Explorative choices lead to new information, with exploitative choices are associated with immediate rewards. Adapted with permission from Mehlhorn et al., 2015.

Behavioral systems

Exploitation refers to the repeated choice of the option with the highest subjective value, while exploration is defined as disengaging from the choice pattern that currently seems optimal, to search for alternative behaviors and acquire new information about the environment (Mehlhorn et al., 2015; Sutton & Barto, 1998). More explorative behavior might be optimal when nothing is known about the environment, whereas exploitation is more optimal when learning has converged to the solution offering maximum reward. Exploration can then again become relevant when a) circumstances change and the option is no longer optimal, or b) potentially better options emerge that were not known before (Figure 5). Research on exploration has suggested that uncertainty plays an important role driving explorative behavior in order to promote the acquisition of new information, possibly through an "information" or "novelty" bonus that is added to unknown options (directed exploration; Kakade & Dayan, 2002; Schulz & Gershman, 2019; Wittmann et al., 2008). Another proposed mechanism is "random exploration", produced by inducing stochasticity into choices (Daw et al., 2006).

The exploration-exploitation tradeoff is often assessed in foraging tasks (Averbeck, 2015; Constantino & Daw, 2015; Mobbs et al., 2018). The setting typically consists of a search for resources in a patchy and changing environment. Participants are tasked to obtain a maximum number of rewards, for example harvesting apples in a virtual orchard (Constantino & Daw, 2015; Lenow et al., 2017). At each trial, they choose between harvesting the current tree or switching to a new tree. At each subsequent harvest, the richness of the tree decreases by a depletion rate. Switching to a new tree on the other hand comes with a travel time during which it is not possible to obtain further rewards (Figure 6). To assess the degree to which participants consider the cost of a travel time in their choices, this time can either be short or long (6 vs. 12 seconds in this example). The task consists of 4 blocks, reflecting different environments, in which the travel times are either short or long, respectively. Exploiting the current resource reduces opportunity costs and uncertainty associated with switching to a new patch but leads to a decreasing value of the resource. Over time, the potential value in exploring a new resource increases, leading to an exploration-exploitation tradeoff. The exit threshold, defined as the average number of apples harvested in the last two trials before leaving to the next tree, therefore indicates whether choices are rather explorative or exploitative. High thresholds (high number of apples left, i.e., early switching) point towards exploration, while low thresholds (late switching) are associated with exploitative behavior.

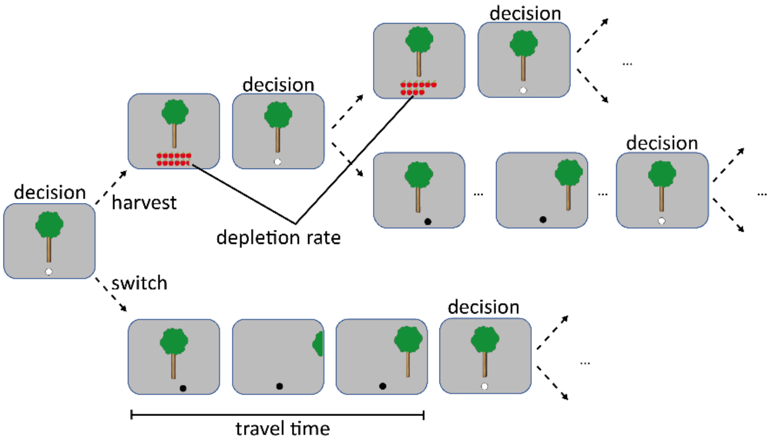


Figure 6. Foraging task. In each trial, participants choose whether to harvest the current tree or to switch to a new tree. The number of apples declines with each subsequent harvest according to a depletion rate. Switching leads to a new, unharvested tree, but comes with the cost of a travel time. Exploration is indicated by fast switching to new options, while exploitation is reflected in staying at one tree, although the reward decreases with every subsequent harvest. Reprinted with permission from Cremer et al. (2023).

Computational Systems

The marginal value theorem (MVT; Charnov, 1976), originally stated in animal literature, addresses how individuals make decisions where they must choose between continuing to exploit a current, depleting resource or moving on to search for a new one (Figure 7, middle).

This model is grounded in the principle that the optimal decision involves comparing the immediate return from exploiting the current resource to the opportunity cost of time, represented by the environment's long-run average reward rate. The optimal policy would be to leave a resource when the expected reward from one more intake falls below the opportunity cost of the time that would be spent acquiring it. Applied to foraging tasks as displayed in Figure 6, the optimal time to switch to a new tree is when the expected number of apples at the next harvest falls below the long-run average reward rate in the current environment (Constantino & Daw, 2015). The expected reward is computed by the reward in the current trial, discounted by the depletion rate. The average reward rate is defined as the average reward rate in the current environment times the time it takes to harvest. By modeling the task structure and entering all possible leaving thresholds, the probabilistically expected rewards over time can be simulated per threshold, ultimately returning the optimal leaving threshold. Individuals staying at the current tree further beyond this threshold, show a bias towards exploitation, while those switching to a new tree way earlier tend to explorative choice behavior. In the MVT learning model (Constantino & Daw, 2015), the average reward rate in the current environment is updated trial-by-trial based on the difference between the actual and the expected reward, namely, the RPE. A learning rate controls the degree to which the RPE leads to an adjustment of action values, and an inverse temperature parameter encodes the extent to which these action values are used to guide decisions. Lastly, an intercept indicates whether behavior reflects a constant choice bias with higher values pointing towards a bias to staying (exploitation) and lower values representing a bias towards switching (exploration).

Neural systems

Substantial experimental evidence suggests that exploitative decisions arise from a corticostriatal network, with the pivotal regions being the vmPFC and the OFC (Figure 7, right). The vmPFC is well known for reward-driven choices (Blanchard & Gershman, 2018; Chakroun et al., 2020; Summerfield & Koechlin, 2008). In line with this, the vmPFC encodes reward anticipation (Kim et al., 2011; Lorenz et al., 2014) and tracks the value of choice options (De Martino et al., 2013; Kolling et al., 2012). The OFC is involved in the subjective valuation of choice alternatives, encoding a map of available options in the current environment (Groman et al., 2019; Stalnaker et al., 2014; Wikenheiser & Schoenbaum, 2016). Exploration on the other side is associated with the frontoparietal control network (Figure 7, right) with the core regions frontopolar cortex (Badre et al., 2012; Beharelle et al., 2015; Boorman et al., 2009; Zajkowski et al., 2017), middle frontal gyrus (Chakroun et al., 2020; Hogeveen et al., 2022), dorsal anterior cingulate cortex (dACC, Blanchard & Gershman, 2018; Kolling et al., 2012) and IPS (Daw et al., 2006; Laureiro-Martínez et al., 2015).

Importantly, there is growing evidence that exploration and exploitation not only rely on distinct neural circuits, but that these processes also involve dopamine and noradrenaline as key neurotransmitters. As described before, choosing options with the highest expected reward is tightly linked to striatal dopamine (see 1.1; Glimcher, 2011; Schultz et al., 1997). Consequently, it stands to reason that exploitation is associated with striatal dopamine transmission (Figure 7, right). This is supported by a finding that genes linked to *striatal* dopamine signaling are also associated with increased exploitation behavior (Frank et al., 2009). At the same time, dopamine has been linked with exploratory behavior, particularly those genes modulating *prefrontal* dopamine function (Figure 7, right). Individuals carrying a variant of the catechol-O-methyltransferase (COMT) gene, associated with elevated tonic dopamine levels, tended to make exploratory decisions in response to uncertainty about whether alternative choices might yield better outcomes than the current situation (Frank et al., 2009). One potential mechanism driving this 'directed' exploration is a novelty bonus, which is applied to unknown alternatives when uncertainty is high to facilitate the acquisition of new information (Zajkowski et al., 2017).

Noradrenaline has also been associated with exploration, but the underlying mechanism differs. Instead of a targeted switch to unknown options, noradrenaline-driven exploration is induced by adding a stochasticity to the decision-making process (‘random exploration’, Dubois et al., 2021). High uncertainty therefore may trigger noradrenaline release, signaling the interruption of ongoing information processing. Often referred to as a ‘reset button’, noradrenaline initiates exploring new options by resetting cached value representations (Dayan & Yu, 2006). In the same vein, rodents with elevated noradrenaline levels showed an increase of value-free-random-like behavior (Tervo et al., 2014), while monkeys showed increased choice consistency after the pharmacological blockade of noradrenaline (Jahn et al., 2018). Another line of research found noradrenaline to be involved in both exploitation and exploration, differentiating phasic and tonic levels. Phasic noradrenaline activity was associated with exploitation, while tonic signaling may facilitate explorative behavior (Aston-Jones & Cohen, 2005; Kane et al., 2017).

Conclusion

The exploration-exploitation framework allows a precise quantification of when to disengage from the current option in favor of a potentially better, but unknown alternative. However, exploration and exploitation are not static, mutually exclusive choices but are dynamically adjusted based on feedback and changing environments. Individuals often engage in a combination of both strategies, adapting their behavior to optimize outcomes. Moreover, they basically share the same mechanisms, as formalized in the MVT learning model. Whether exploration and exploitation arise from two different systems or rather represent two poles of a continuum cannot be fully determined from the current state of literature.

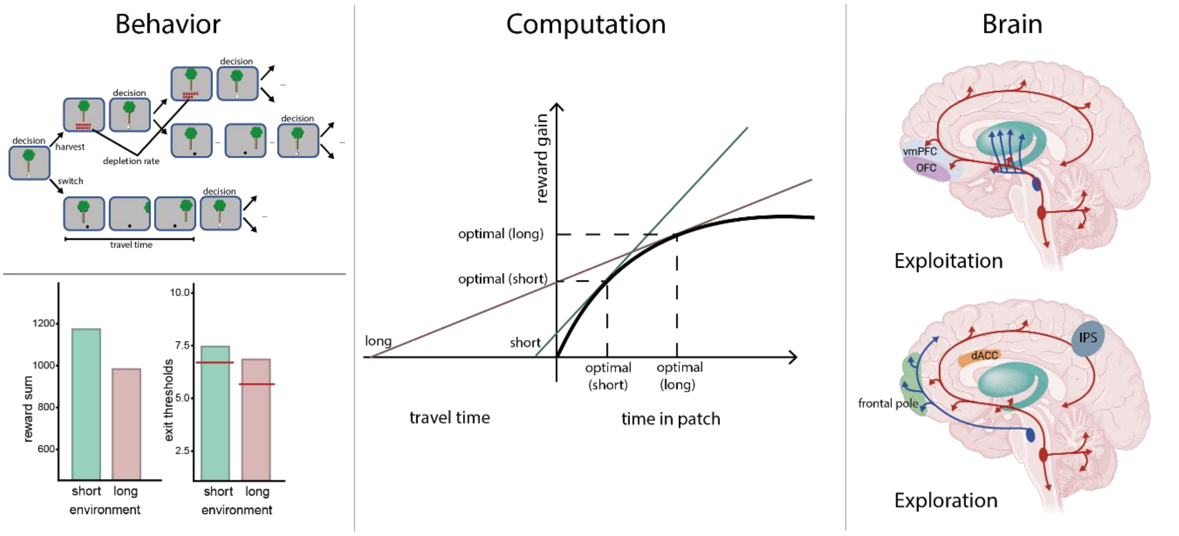


Figure 7. Summary of the exploration-exploitation dichotomy. Left: Explorative vs. exploitative choice behavior is typically assessed in foraging tasks where participants choose between exploiting the current resource with decreased returns at each subsequent harvest or switching to a new resource with the cost of a travel time (top). Staying longer at the current option than the optimal leaving threshold indicated a bias towards exploitation, while earlier exits reflect a bias towards exploration (bottom). Reprinted with permission from Cremer et al., 2023. Middle: The MVT states that the optimal time to switch to a new tree is when the expected number of apples at the next harvest falls below the long-run average reward rate in the current environment. Adapted from Halsey & Butler, 2006. Right: Key regions associated with exploitation are the vmPFC, and the OFC. Striatal dopamine transmission (blue arrows) is linked to exploitative choices (top). Exploration is associated with the frontoparietal network, including the frontopolar cortex, the dACC, and the IPS. Prefrontal dopamine (blue arrows) is implicated in explorative decisions (bottom). Noradrenaline (red arrows) was associated with both exploitation and exploration.

1.2.3 Reward learning vs. working memory

As described before, typical RL approaches focus on the process of associating state-action pairs with future outcomes to find policies that maximize rewards. However, this learning process critically depends on cognitive resources that allow an agent to store transient value information. Specifically, RL requires the active maintenance of a set of stimulus-action-outcome associations that are constantly updated during learning. WM is a memory system well-suited to these requirements (Baddeley, 1992). Both RL and WM are central domains of human cognition but are mostly considered distinct processes – both at the behavioral and the neural level (Yoo & Collins, 2022). However, they may be less dissociable than often assumed. Collins and Frank (2012) proposed an interplay between reward learning and a capacity-limited, decay-sensitive WM process that co-determines learning, but can be disentangled. In the context of RL, WM's primary role is to hold representations of behaviorally relevant stimulus-action-associations in an accessible state, serving as basis for immediate decisions (Baddeley, 1992; Durstewitz et al., 2000; Miller et al., 2018). A classic example is remembering a shopping list and recalling the items as walking through the store. Two properties of this task can be changed to make it more taxing on WM. First, the sequence length: The longer the list, the more difficult it is to remember all entries. Second, the retention interval: The more time elapses between creating the list and applying it in the store, the more challenging it is recalling the full list. These limitations have been widely confirmed in WM literature (Durstewitz et al., 2000; Miller et al., 2018). Experimental tasks typically involve holding a set of representations in mind over a delay and then being tested on the recall accuracy. Thereby, the length of the delay and the number of memoranda are being manipulated. With either an increasing number of representations or longer retention intervals, the recall accuracy decreases (Cowan, 2001, 2008; Oberauer et al., 2016).

Behavioral systems

WM is essential in any choice situation, as it makes the options available by maintaining them in a temporal space. Therefore, the question is not whether RL and WM are interacting, but rather how. This question has been directly addressed in a stimulus-association task developed by Collins & Frank (2012). In a stimulus-response association task, participants learned the correct action for a varying number of stimuli within one set. Correct answers were always rewarded with one point but could probabilistically lead to an increased reward of 2 points. RL demands were varied through the probability of receiving an extra reward. Accumulated rewards reflected the Q-value for each stimulus. To address capacity- and time-sensitive contributions of WM, the number of stimuli within one set (*load*) and the number of trials between stimulus repetitions (*delay*) were manipulated systematically (Figure 8A, learning phase). Results showed that both RL and WM determined learning: The number of previous correct decisions, reflecting RL-related cumulative value, increased the probability to choose the correct action in subsequent trials. At the same time, performance decreased with increasing delay and set size, reflecting the influence of WM limiting task performance. In surprise test phases, the contributions of WM and RL were further disentangled.

In a first test phase, participants were presented with two stimuli from the learning phase and had to indicate for which stimulus they had obtained more rewards in the learning phase. Stimuli were mixed across blocks, addressing the cumulated Q-value computed by the RL-system for each stimulus (Figure 8B; Collins, Albrecht, et al., 2017). Interestingly, participants were more sensitive towards value differences between the two stimuli when the stimuli came from blocks with higher set sizes, indicating that Q-value representations were more robust when learning was predominantly driven by RL (Collins, Albrecht, et al., 2017).

In a second test phase, the retention of the stimulus-response associations was probed (Figure 8C; Collins, 2018). Randomly intermixed stimuli from the learning phase were presented and participants should indicate the correct action associated with the stimuli without receiving feedback. This phase took place at least 10 minutes after the last block of the learning phase when WM representations should no longer be accessible. Associations learned under low load were retained worse than those learned under high load, supporting previous findings that fast WM learning comes with the cost of higher decay, while slow and effortful RL is more robust over time (Collins, 2018).

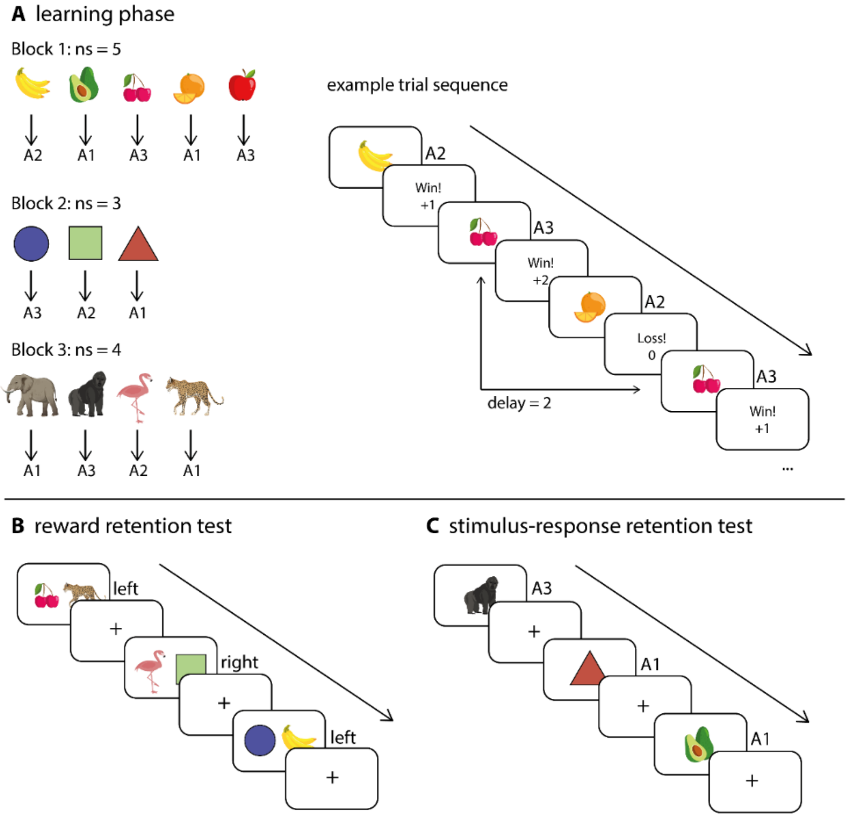


Figure 8. A stimulus association task to separate RL and WM contributions to learning. (A) Learning phase. In each trial, participants were presented with a stimulus and had to choose one out of three actions, deterministically associated with a reward. RL components were addressed by a varying amount of reward associated with a stimulus: Correct answers could lead to an increased reward of 2 points with a fixed probability of 0.2, 0.5, or 0.8. WM contributions to learning were determined by the number of stimuli within a block (set size) for which the associated action had to be learned and by the delay between the last correct response to a stimulus and its subsequent presentation. (B) Reward retention test. In a surprise test phase addressing the cumulative reward representation, participants were asked to indicate for which out of two stimuli from the learning phase they obtained more rewards. No feedback was given. (C) Stimulus-response-retention test. In a second surprise test phase, participants were presented with stimuli from the learning phase and had to choose the action that was associated with a reward without getting feedback. Reprinted with permission from Rac-Lubashevsky et al. (2023).

Computational Systems

The disentanglement of WM and reward learning processes was further supported by a computational model that combined a model-free RL process and a time-sensitive, capacity-limited WM process (Collins, 2018; Collins, Albrecht, et al., 2017; Collins, Ciullo, et al., 2017; Collins et al., 2014; Collins & Frank, 2012, 2018; McDougle & Collins, 2021). In this reinforcement learning working memory (RLWM) model, the classic RL module learns incrementally from RPEs, adjusted by a learning rate α that controls the extent to which the value of an option is updated by the RPE. As described before, it further contains an inverse temperature parameter β determining the degree to which differences in Q values lead to the adaption of choices. The WM component can learn immediately but is capacity-limited and decaying over time. A decay parameter accounts for potential forgetting by pulling Q value estimations towards their initial value, and a capacity parameter limits the probability of WM usage in higher set sizes. Which system guides choices is determined by the weighted average of the policies of the two modules. In turn, this weighted average depends on the probability that a stimulus was stored in WM, as a function of the set size. Support for this RLWM model came from the observation that it provided a better fit to the data than the separate modules alone.

Neural systems

WM is supported by a network of neural mechanisms involving various brain regions and processes. Early findings characterized persistent firing in the dlPFC, specifically in the medial frontal gyrus (MFG) and in the superior frontal gyrus (SFG), as the core neural substrate of WM (Cohen et al., 1997; Funahashi et al., 1989; Fuster, 1973; Goldman-Rakic, 1995). Since then, it has been shown that the WM processes in the prefrontal cortex (PFC) might also represent the maintenance of higher order information, such as rules, goals, or abstract representations of categories relevant to behavior (Lee & D'Esposito, 2012; Riggall & Postle, 2012; Sreenivasan et al., 2014). According to this rules or goals, top-down signals from the PFC bias the salience of mnemonic representations, which influence the entry of information into WM storage within the PPC (Berryhill et al., 2011; Murray et al., 2017). More specifically, dopaminergic interactions between the PFC and the basal ganglia (1) determine which stimuli pass the gate to WM, (2) prevent maintained representations from distracting information, (3) allow relevant representations to update and (4) select which information from WM is relevant for behavior (Chatham et al., 2014; Cools & D'Esposito, 2011; D'Esposito & Postle, 2015; Gruber et al., 2006; Hazy et al., 2006; Ott & Nieder, 2019; Rac-Lubashevsky & Frank, 2021). In summary, WM functioning emerges from a complex interplay between PFC, PPC, and striatal circuits (Figure 9, right). As described before, these regions are also involved in both the model-based vs. model-free (Figure 4, right) and in the exploration-exploitation framework (Figure 7, right), which highlights the need for a deeper understanding of WM and RL contributions to learning. Neuroimaging studies aiming to disentangle WM an RL confirmed set size effects in the SFG and RPE activity in nucleus caudatus and thalamus (Collins, Ciullo, et al., 2017; Collins & Frank, 2018). Interestingly, the neural signature of RPEs was modulated by set size: RPE signals were weaker in lower set sizes in which subjects' learning was closest to optimal, and thus likely to be driven by WM. At the same time, RPE signals were stronger under high load, accompanied by enhanced value learning, indicating that the interplay between WM and RL is not modular, but rather interactive (Collins, Ciullo, et al., 2017; Collins & Frank, 2018).

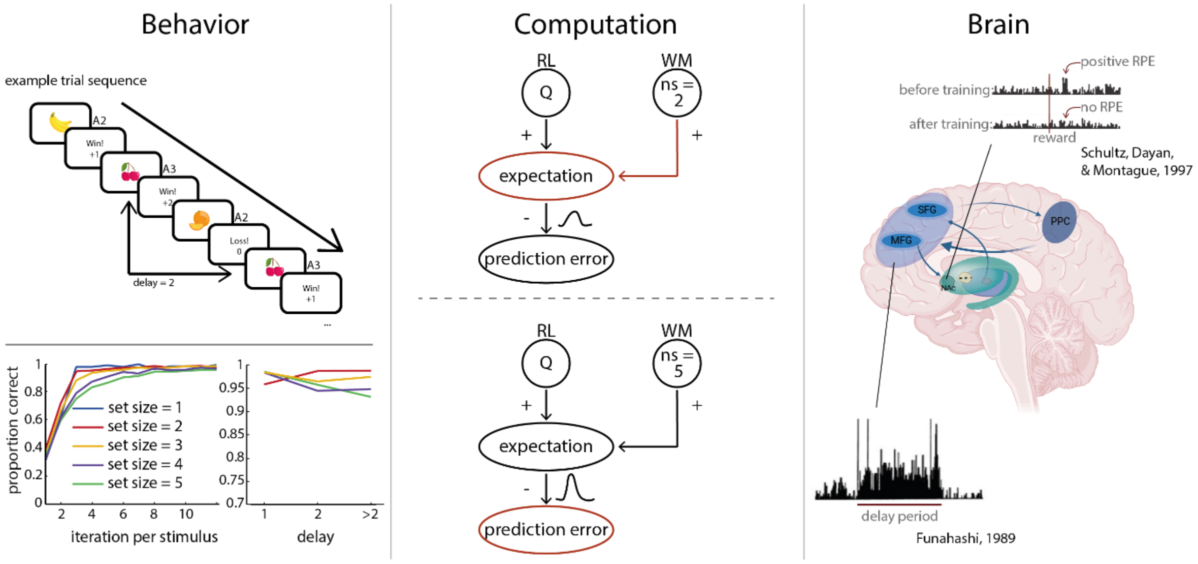


Figure 9. Summary of the WM-RL dichotomy. Left: Disentangling of WM and RL behavioral signatures in a stimulus response task. RL components are addressed by varying amounts of rewards associated with stimulus-action pairs, while WM processes are determined by the number of stimuli within one set and the number of trials since the current stimulus was last responded to (top). WM enables fast learning for smaller sizes, while a slower, but broader RL process takes over for larger sizes (bottom, left). Performance decreases with increasing delay between stimuli, especially in higher set sizes (bottom, right). Middle: The RLWM model uses RPEs to update value estimates (Q) of state-action pairs actions to guide future decisions. Both systems give input to Q: When the number of stimuli is within WM capacity, Q-value estimates are informed by WM, leading to fast learning and reduced RPEs (top). In trials with higher load, WM capacity is exceeded, resulting in slower learning and higher RPEs (bottom). Reprinted with permission from Rac-Lubashevsky et al., 2023. Right: RL and WM rely on overlapping brain networks, both modulated by dopamine. RPE computations are tightly liked to dopaminergic firing in the ventral striatum (top inset). Key regions associated with WM are the dlPFC (including MFG and SFG) and the PPC. Prefrontal dopaminergic firing signals maintaining information in WM in a delay-period (bottom inset). Adapted with permission from Yoo & Collins (2022).

Conclusion

While the dual process model of WM and RL provides a sophisticated framework for understanding the interaction between these two core cognitive processes, some critical points and challenges are worth considering. The dual-systems model simplifies the neural bases of WM and RL, often highlighting distinct regions such as the prefrontal cortex for WM and the basal ganglia for RL. However, neural mechanisms are highly interconnected, with significant overlap and shared pathways that the model may not fully account for. This complexity suggests that WM and RL may not be as separable in their neural substrates as the model implies. Although the approach proclaims an integrative perspective, it also involves the detached consideration of two processes. The isolated focuses on WM and RL does not incorporate the roles of other cognitive processes such as attention, executive function, and long-term memory. These processes undoubtedly influence and are influenced by WM and RL, suggesting that an even more integrative model might be necessary.

1.3 Modulation by stress and stress mediators

Stress is a powerful modulator of learning and decision-making (Luksys & Sandi, 2011; Porcelli & Delgado, 2017). The effect of stress on learning and decision-making processes is an integral part of the adaptive fight-or-flight response, helping to effectively cope with the stressor (McEwen, 1998; Ulrich-Lai & Herman, 2009). In the response to acute stress, a well-coordinated cascade of different physiological and endocrine changes is initiated, which is essentially determined by two pathways (McEwen, 1998; McEwen, 2007). When exposed to stress, the *autonomic nervous system* is immediately activated, followed by the slower reaction of the *hypothalamic-pituitary-adrenal axis* (Joëls & Baram, 2009). Seconds after stressor onset, the autonomic nervous system triggers the release of catecholamines such as adrenaline, noradrenaline, and dopamine to create alertness and provide energy to prepare the organism for action (De Kloet et al., 2005; Joëls & Baram, 2009). The second pathway, the hypothalamic-pituitary-adrenal axis, is activated somewhat more slowly than the autonomic nervous system triggering the synthesis and release of corticosteroids (mainly cortisol in humans) into the bloodstream (De Kloet et al., 2005; Joëls & Baram, 2009). Cortisol levels peak around 20-30 minutes after stressor onset and bind to receptors that are widely expressed in regions implicated in RL and decision-making, such as the PFC, the amygdala and the hippocampus (Hermans et al., 2014; Joëls & Baram, 2009).

Following acute stress, cortisol binds to glucocorticoid receptors in the PFC, disrupting synaptic plasticity and dendritic remodeling, which are critical for RL (Holmes & Wellman, 2009). Further, elevated levels of catecholamines, especially noradrenaline, lead to the activation of α1-adrenergic receptors and the inhibition of α2-adrenergic receptors in the PFC. This shifts the PFC from a state of executive control to a more reactive and less organized state, impairing its function (Arnsten, 2009). Another immediate effect of elevated glucocorticoid levels is the suppression of long-term potentiation in the hippocampus, resulting in reduced synaptic plasticity, thereby impairing the hippocampus' ability to encode and consolidate new memories (Kim & Diamond, 2002).

Within the context of dual-systems approaches, stress was found to induce a shift from computationally demanding systems to rather rigid, reactive behavioral strategies (Schwabe & Wolf, 2011; Wirz et al., 2018). Higher cortisol levels were found to selectively reduce model-based contributions to behavior while model-free learning processes stayed intact (Otto et al., 2013). At the same time, stress was associated with an increase in striatal, not prefrontal, dopamine (Anstrom & Woodward, 2005), contributing to the enhanced reliance on model-free learning (Park et al., 2017). Further, the stress-induced noradrenaline release in the PFC was found to induce a shift from phasic to tonic mode of noradrenaline firing that led to less focused attention to specific stimuli, replaced by a "scanning-mode" of the environment, thus promoting explorative behavior (Aston-Jones & Cohen, 2005; Hermans et al., 2014; Rajkowski et al., 1997). Conversely, other studies suggest a behavioral shift in the direction of exploitation, showing that stressed participants were more likely to rely on familiar strategies and known rewards rather than exploring new options. (Lenow et al., 2017; Luksys & Sandi, 2011). Finally, the detrimental stress effects on PFC functioning also impact WM. Both the immediately elevated noradrenaline levels and the subsequent release of cortisol were associated with a decrease in prefrontal delay cell activity, leading to an impaired WM functioning (Birnbaum et al., 2004; Ramos et al., 2005).

1.4 Research goals

Dual-systems approaches have a long tradition in psychology and neuroscience and are often the source of groundbreaking theories (Evans & Stanovich, 2013). At the same time, they have been criticized for being oversimplistic, arguing that a binary classification is not suitable to capture the complex mechanisms underlying learning and decision-making (Gigerenzer, 2010; Kruglanski & Gigerenzer, 2011). In the field of RL, two-systems approaches are applied in various contexts such as model-based vs. model-free RL, exploration vs. exploitation, and WM vs. reward learning. While these frameworks have been studied extensively detached from another, an integrative perspective is lacking. Considering the abundant overlaps in computational components and neural regions involved, the integration of findings across frameworks can add significant value in understanding the mechanisms underlying value-based decision-making. While some parts of the RL process are beyond doubt, such as the dopaminergic signaling of RPEs in the striatum (Bayer & Glimcher, 2005; Schultz et al., 1997), the exact identification of the systems' pathways and their interplay are widely unknown.

We set out to investigate the existence of dissociable systems and their susceptibility to a modulation by stress and stress mediators. Through a combination of behavioral and neural data, the aim was to explore the extent to which dual processes are distinguishable, and to which extent the use of one vs. the other strategy can be modulated experimentally. We tested these questions in the context of decision-making tasks, where different patterns of choice behavior correspond to distinct cognitive strategies.

In Study 1, we focused on comparing model-based and model-free learning in a two-step task, specifically under stress. We modulated the two-step task by introducing reversals of reward contingencies to promote the use of model-based learning. The main hypothesis proposed that inducing stress should lead to a shift towards a more model-free learning strategy. This shift should be reflected in strategy-specific neural activation patterns across different regions, which we assessed using functional magnetic resonance imaging (fMRI). Finally, hypotheses were formalized through computational modeling. Overall, Study 1 aimed to uncover the impact of stress on the interplay between model-based and model-free learning strategies, and to provide insights into associated neural mechanisms.

Study 2 aimed to test the putative dichotomy of exploration vs. exploitation strategies in a foraging task. It sought to determine whether these strategies indeed represent distinct systems by selectively blocking neurotransmitters associated with each, namely dopamine and noradrenaline. Moreover, we used computational modeling to measure the impact of this pharmacological manipulation on the preferred use of exploration- vs. exploitation-based policies. This study aimed to find neurochemical evidence for distinguishable RL systems guiding decision-making in foraging tasks.

Study 3 assessed to which extent WM and RL are alternative systems, with a specific focus on the impact of stress. We expected that stress would predominantly influence prefrontal executive functions. Assuming a strict dichotomy between WM and RL, stress should selectively disturb WM-depended learning, while leaving RL-depended learning largely unaffected. The study aimed to identify WM and RL signatures on the behavioral, computational, and neural level. We employed electroencephalography (EEG) to explore neural correlates of WM and reward learning, and tested whether they were affected by the stress manipulation. In summary, Study

3 employed a stress manipulation to uncover possible behavioral, computational and EEG signatures that are unique indicators of WM- vs. RL-based learning.

Together, these studies investigated whether dichotomized systems in RL, as proposed in the literature, can be experimentally tracked, and differentiated. We further tested whether the systems, their interplay, or specific components were subject to the effects of stress and stress mediators. Computational modeling was used to track such behavioral changes in fitted parameters of mathematical learning models. Finally, EEG and fMRI recordings allowed us to assess whether these systems are reflected in separable patterns of neural activity.

2 Experimental Studies

2.1 Study 1: Does stress induce a shift from model-based to model-free RL contributions to choice behavior?

Cremer, A., Kalbe, F., Gläscher, J., & Schwabe, L. (2021). Stress reduces both model-based and model-free neural computations during flexible learning. *NeuroImage*, *229*, 117747. https://doi.org/10.1016/j.neuroimage.2021.117747 – (Appendix A)

2.1.1 Background

Stress is a powerful modulator of learning and decision-making (Luksys & Sandi, 2011; Porcelli & Delgado, 2017; Schwabe et al., 2011; Schwabe & Wolf, 2009; Shafiei et al., 2012). The stress-induced release of monoamines and glucocorticoids, in particular, impairs prefrontal function (Arnsten, 2009; Cerqueira et al., 2007; Ossewaarde et al., 2011; Vogel et al., 2016). This suggests that stress might primarily affect model-based learning, rather than model-free learning. Therefore, we tested the presence of these separate systems underlying RL and how their interplay was modulated by stress. Preliminary studies indicate that acute stress induces a shift from model-based to model-free behavior (Otto et al., 2013; Park et al., 2017; Raio et al., 2017), but the underlying mechanisms are widely unknown. To address this issue, we tested behavioral and neural signatures of model-based and model-free learning systems during choice behavior, and how these were affected by stress. Specifically, we quantified (i) the relative contribution of model-based and model-free learning to decision-making, (ii) whether the strategies were associated with separable neural systems, and (iii) whether stress induced a shift from the model-based to the model-free system.

2.1.2 Methods

Fifty-eight healthy volunteers underwent either the Trier social stress test (Kirschbaum et al., 1993) or a control manipulation, before performing a modified version of a two-step task in an MRI scanner in which reward contingencies in the second stage were reversed at several random time points throughout the task. We hypothesized that this manipulation should promote model-based RL, since the internal task model that defines model-based RL should facilitate the adjustment to changing reward probabilities (Akam et al., 2015). This would enable us to pinpoint the hypothesized stress-related shift from model-based to model-free learning.

2.1.3 Results

Our behavioral results indicate that subjects used mainly model-free learning, as they repeated rewarded actions regardless of the transition type. At the same time, the contribution of model-based RL, as indicated by a reward × transition interaction, was also reflected in our data (Figure 10). However, computational modeling results show an overall bias towards model-free RL (weighting parameter w < 0.5 for both stress and control groups, where w = 1 reflects pure model-based RL, while w = 0 stands for pure model-free RL). Interestingly, the temperature parameter β for the first stage choice tended to be lower in the stress group compared to the control group, pointing towards a more random first stage choice behavior in stressed participants.

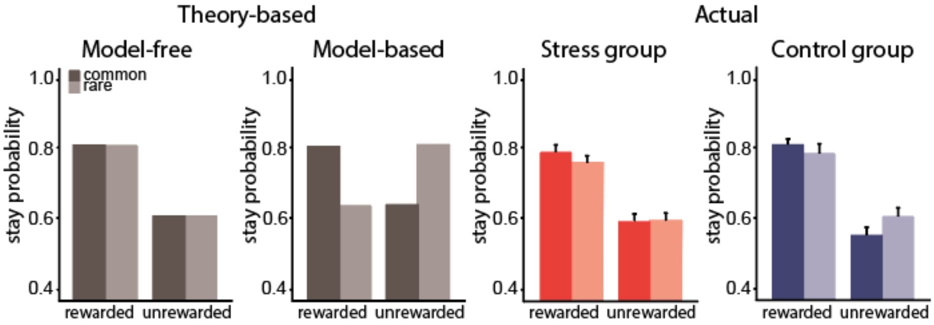


Figure 10. Factorial analysis of choice behavior. Left: Pure model-free reinforcement learning predicts that a previously rewarded action is more likely to be repeated on the subsequent trial, regardless of whether the transition was common or rare. Pure model-based behavior considers the task structure: a reward obtained after a rare transition should promote a switch to the other option. Right: Actual Data. Participants show both model-based and model-free learning with an overall bias toward model-free learning. There were no significant group differences. Adapted with permission from Cremer et al. (2021).

On the neural level, stress reduced contributions of both model-based and model-free computations. Model-free PEs in the stress group were paralleled by a decrease of ilPFC activity in comparison to the control group (Figure 11A). Model-based PEs were also associated with a stress-induced activity reduction, located in the posterior hippocampus and in the putamen. At the same time, the stress group showed increased activity in the ilPFC during model-based PE computation, compared to the control group (Figure 11B).

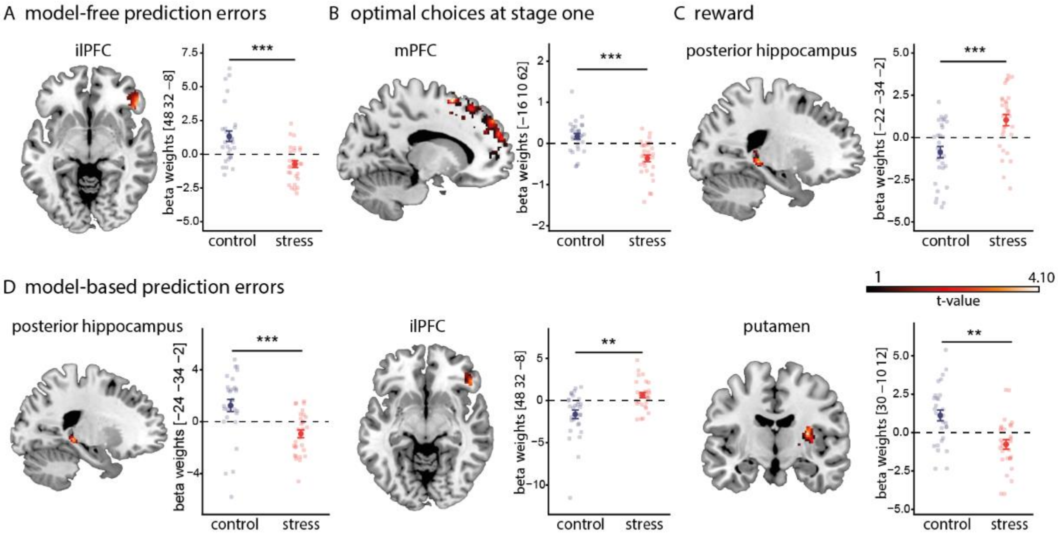


Figure 11. Stress effects on the neural mechanisms underlying model-based and model-free RL. Stress did not specifically hamper model-based RL, but reduced error computations in both model-based and model-free RL. Activity in the ilPFC was higher during model-based prediction errors in the stress group, compared to the control group. Reprinted with permission from Cremer et al. (2021).

2.1.4 Conclusions

Our behavioral results show that participants used a mixture of model-based and model-free RL. Although our task modification of frequently reversing reward contingencies should favor a model-based strategy, participants in both groups showed an overall bias towards model-free RL. At the same time, we hypothesized that stress would hamper processing in the PFC, accompanied by a reduced use of a model-based learning strategy. We identified group differences in neural activity in the ilPFC both during model-based and model-free prediction error computation. Differential activation in the experimental groups were further present in the posterior hippocampus, a) at reward signaling and b) during model-based prediction errors. In summary, our behavioral, computational, and neuroimaging results do not support a clear separation of model-based and model-free RL systems.

2.2 Study 2: How do dopamine and noradrenaline influence the exploration-exploitation tradeoff?

Cremer, A., Kalbe, F., Müller, J. C., Wiedemann, K., & Schwabe, L. (2023). Disentangling the roles of dopamine and noradrenaline in the exploration-exploitation tradeoff during human decision-making. Neuropsychopharmacology, 48(7), 1078-1086. https://doi.org/10.1038/s41386-022-01517-9 – (Appendix B)

2.2.1 Background

Knowing when to leave an option for a potentially better, but unknown alternative is at the heart of the exploration-exploitation-dilemma. Both dopamine and noradrenaline have been suggested to play a role in balancing exploration and exploitation (Chakroun et al., 2020; Cohen et al., 2007; Mehlhorn et al., 2015). Paradoxically, both noradrenergic and dopaminergic firing have been linked to exploration (Dubois et al., 2021; Frank et al., 2009; Gershman & Tzovaras, 2018), while other studies suggest a role of noradrenaline in exploitation (Aston-Jones & Cohen, 2005). These seemingly conflicting findings indicate that neither transmitter can be linked to explorative vs. exploitative behavior per se, but that they may signal specific choice aspects. We identified the initial value of an option, the decline of this value over time, and the opportunity cost associated with the switch to another option as key variables in deciding whether to stick with the current option or to explore something new. In this study, we aimed to disentangle the dopaminergic and noradrenergic contributions in processing specific choice aspects underlying the exploration/exploitation tradeoff.

2.2.2 Methods

Sixty-nine volunteers received 400mg of the dopamine D2/D3 receptor antagonist amisulpride, 40mg of the β-adrenoceptor antagonist propranolol, or a placebo, before they performed four blocks of a patch-foraging task. Participants had to harvest virtual apple trees, aiming to yield as many apples as possible within 7 minutes per block. In each trial, they had to decide whether to harvest the current tree, or to switch to a new one. At each subsequent harvest of a tree the reward decreased by a *depletion rate*, which was fixed per tree. Consequently, at a certain point it was advantageous to leave the current tree. Switching, however, came with the cost of a *travel time* that could either be short (6 seconds) or long (12 seconds). We assessed how (i) the different choice components influenced whether to stay at the current tree or to switch to the next option and (ii) how these effects were modulated by the pharmacological manipulation in a mixed-effects logistic regression. Specifically, we were interested in the differential processing of the reward value, the reward decline, and the switching costs.

2.2.3 Results

Participants in the amisulpride group showed greater sensitivity to all three choice features, compared to the placebo group. Specifically, they switched more when either the travel time was short, when the depletion rate was high, or when the number of previous rewards was low (Figure 12). This points towards an increased susceptibility to choice-relevant information after amisulpride intake. Interestingly, participants in the propranolol group tended to show a reduced use of value information, as indicated by the tendency to switch more after high rewards, compared to the placebo group (Figure 12A).

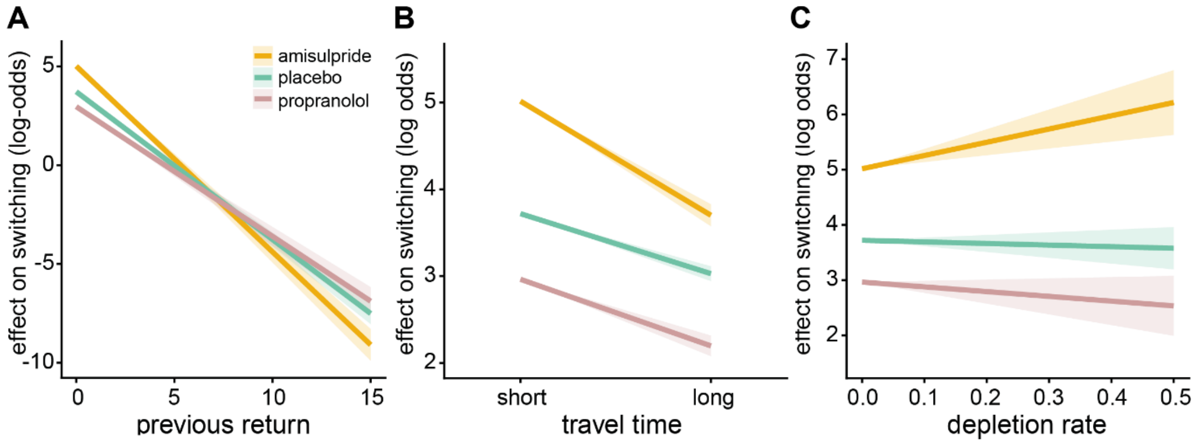


Figure 12. Effects of amisulpride and propranolol on the impact choice features on switch decisions. (A) Participants switched less after high rewards in the amisulpride group, while participants in the propranolol group tended to switch more when the previous reward was high (both compared to placebo). (B) Participants in the amisulpride group switched more when the travel time was short, compared to placebo, while the propranolol group did not differ from placebo. (C) Amisulpride intake was linked to more switching at high depletion rates, while there was no difference between participants in the propranolol and placebo groups. Reprinted with permission from Cremer et al. (2022).

2.2.4 Conclusions

Our results revealed functionally dissociable roles of dopamine and noradrenaline in the processes underlying explorative and exploitative choice strategies. Specifically, our data support the assumption that dopamine is involved in signaling choice-relevant information, while noradrenaline could not be directly linked to the processing of specific choice features. Taken together, the results indicate a role of dopamine in the processing of behaviorally relevant information, while noradrenaline may exert higher order control signals akin to a reset button. The results highlight the importance of considering the specific mechanisms through which these neurotransmitters exert their effects, suggesting that exploration and exploitation decisions are influenced by a range of neurobiological processes. Thus, a critical view might question whether a dichotomous framework can adequately account for the subtleties and variations in how these decisions are made and modulated.

2.3 Study 3: Which neural mechanisms orchestrate the interplay between working memory and reinforcement learning?

Rac-Lubashevsky, R., Cremer, A., Collins, A. G., Frank, M. J., & Schwabe, L. (2023). Neural Index of Reinforcement Learning Predicts Improved Stimulus–Response Retention under High Working Memory Load. *Journal of Neuroscience*, *43*(17), 3131-3143. https://doi.org/10.1523/JNEUROSCI.1274-22.2023. – (Appendix C)

2.3.1 Background

The successful learning of stimulus-action-outcome associations builds on joint contributions of at least two systems, namely WM and RL. While WM enables fast learning and maintaining of behaviorally relevant stimulus-action-associations, RL provides a slower, but broader learning process that is less restricted by memory capacity and time (e.g., Collins, Albrecht, et al., 2017; Collins & Frank, 2012, 2018). Learning was found to dynamically shift from capacity-sensitive WM to RL when the stimulus load is high (e.g., Collins & Frank, 2018), and (ii) to shift from delay-sensitive WM to RL over the course of learning (e.g., Collins, Albrecht, et al., 2017). Associations learned under high WM load were acquired more slowly but the retention was more robust and reflected in larger neural indices of RL (Collins, Ciullo, et al., 2017; Collins & Frank, 2018). In this study, we tested the contributions of both systems to successful learning by parametrically manipulating WM load and delay in a learning task designed to disentangle WM and RL. As in Study 1, we used acute stress as modulating factor to further differentiate the systems. With the stress-induced impairment in the PFC (Arnsten, 2009; Cerqueira et al., 2007; Ossewaarde et al., 2011; Vogel et al., 2016), we hypothesized that stress would specifically reduce WM-guided learning by weakening the relative contribution of WM in the learning phase. We performed two surprise memory tests to further track indicators of WM- vs. RL-based learning strategies. Finally, we tested whether high WM load was associated with increased RPEs and whether this was predictive of retention performance using single-trial decoding of EEG-signatures.

2.3.2 Methods

Eighty-six healthy volunteers underwent either the socially evaluated cold pressure test (SECPT, Schwabe & Schächinger, 2018) or a control manipulation before they performed a stimulus-response association learning task while being recorded by EEG. The task consisted of a learning phase and two surprise test phases. We tested the contributions of both systems to successful learning by parametrically manipulating WM load and delay. In each trial, participants were presented with a stimulus and had to select one out of three actions, deterministically associated with a reward. RL components were manipulated by introducing an additional probabilistic bonus reward for correct responses. Importantly, the probabilities for this bonus reward were stimulus-linked with values of 0.2, 0.5, or 0.8. WM contributions to learning were manipulated by the number of stimuli within a block (set size) for which the associated action had to be learned and by the delay between the last correct response to a stimulus and the subsequent display. In a *reward retention test* (test phase 1), participants were presented with two stimuli from the learning phase, randomly paired across blocks, and had to choose the stimulus they had collected more rewards for. In the *stimulus-response retention test* (test phase 2), participants were again presented with stimuli from the learning phase and asked to indicate which action was associated with this stimulus. No feedback was given at either phase.

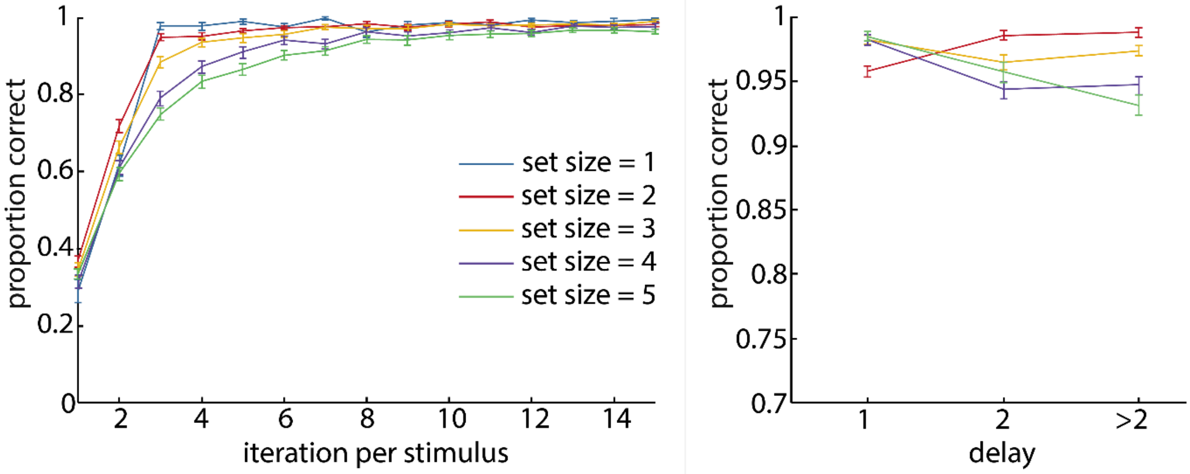


Figure 15. Proportion of correct answers for each set size as a function of stimulus iterations and delay. Left: Correct answers were learned faster for low set sizes, the correct stimulus-action pairs were learned gradually slower with increasing set sizes. Right: the proportion of correct answers decreased with increasing delays in larger set sizes. Reprinted with permission from Rac-Lubashevsky et al. (2023).

2.3.3 Results

Participants showed both WM and RL contributions in the learning phase, indicated by behavioral, computational, and EEG results. RL was reflected in an increasing proportion of correct answers over the course of a block, as expected for incremental value learning. WM-guided learning was less gradual, therefore faster and with higher proportions of correct answers in lower set sizes and a lower proportion of correct answers with increasing delays in larger set sizes (Figure 13). The stress group showed a more pronounced performance drop in blocks with high set sizes compared to the control group, but did not attenuate WM per se. In the reward retention test, participants were more likely to select the item that was more rewarded, especially when the reward values were learned in higher set sizes. Additionally, the stimulus-response retention was parametrically better for associations learned in higher set sizes. Strikingly, during learning, performance was parametrically worse for associations learned in higher set sizes.

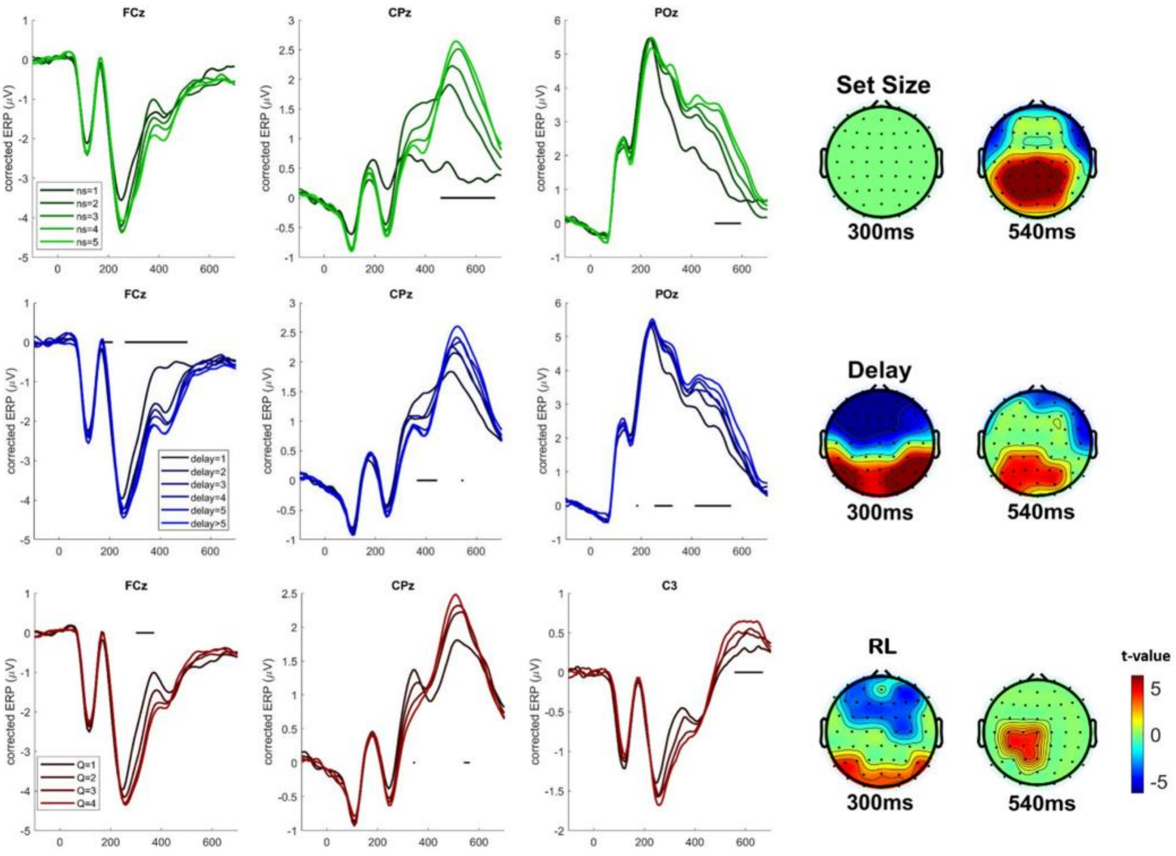


Figure 17. EEG decoding of WM and RL contributions in the learning phase. Corrected event-related potentials (ERPs) confirmed significant effects for both WM and RL contributions. The WM-related predictors *set size* (top row) and *delay* (middle row) were associated with significant activity in frontal and parietal regions (electrodes FCz, CPz, and Poz). Frontal and parietal *delay* effects were initiated 300ms after stimulus onset, reaching a second peak ~540ms, when also the parietal *set size* effect peaked. RL-driven *Q-values* (bottom-row) exhibited significant activity for the electrodes FCz, CPz, and C3, starting with an early frontal activity (~300 ms after stimulus onset) and a second temporal activity peak (~600ms after stimulus onset). Reprinted with permission from Rac-Lubashevsky et al. (2023).

EEG analyses confirmed that RPE signals increased faster in trials with high set sizes than in trials with low set sizes, although behavioral learning was slower in these conditions (Figure 14). Most importantly, these neural signatures predicted better retention of learned associations over time: neural indices of RL during acquisition were predictive of the retention of stimulus-response associations, but not predictive of the stimulus-reward value retention. Our results further indicated that although stressed participants showed a greater performance drop in higher set sizes during leaning, stress did not lead to a decrease of WM performance per se.

2.3.4 Conclusions

Our results confirm a cooperative interplay between WM and RL with the advantages of fast learning via WM at low load, and a robust retention for stimuli learned by the RL module at high load. The results from the learning task demonstrate that associations in blocks with high load were learned slower but showed higher retention accuracy in the end, supported by EEG signals of RL components that increased more rapidly across trials under high than low load, even though behavioral learning was slower in these conditions. Importantly, neural indices of RL during acquisition were predictive of retention of S-R association but not predictive of reward value retention. Our results support the key model prediction that RL processes play a major role in enhancing policy retention when WM resources are depleted.

3 General Discussion

In cognitive science, learning and decision-making behavior has been dichotomized across various dimensions. The theoretical formalization is usually based on the assumption of one system that acts fast, reactively, implicitly and retrospectively, while the opposite system acts slowly, deliberatively, explicitly and prospectively (Collins & Cockburn, 2020). We investigated the extent to which the dichotomies suggested by theoretical RL frameworks can be confirmed in the experimental context. Specifically, we aimed to test the notion that two systems, dissociable in behavioral, computational, and neural mechanisms, drive learning and decision-making in three different contexts. We addressed the dissociation of model-free and model-based RL (Study 1), exploration and exploitation (Study 2), as well as the interplay between WM and reward learning (Study 3). Our objective was to investigate the distinguishability of each of these dual-systems frameworks and to examine the degree to which stress or stress mediators can influence the preference for one strategy over the other.

3.1 Stress-induced reduction of both model-based and model-free signatures in the brain

The dissociation of model-based and model-free behavior is one of the most influential dual-systems framework in the field of RL. Our data show that participants used both model-based and model-free learning strategies in the modified two-step task with an overall bias towards model-free learning both in the stress group and in the control group. A possible explanation is that the modification of adding reversals of reward contingencies at random points throughout the task increased task difficulty and made it challenging to build an internal model of the task structure, therefore biasing participants towards a model-free choice behavior. This was further supported by computational modeling results, which failed to provide evidence that acute stress specifically hampered the model-based learning signatures commonly proposed in the literature (Park et al., 2017). Despite the absence of a stress effect specific to model-based learning, our computational modeling results still showed behavioral differences between the stress group and the control group. Our results revealed that the inverse temperature parameter β for choices in the first stage tended to be lower in the stress group, compared to the control group. The inverse temperature parameter indicates the degree to which value estimates influence choice with lower values reflecting less use of value computations, leading to more random choices. Consequently, stress might have led to a behavior in which the decisions in the first stage were not used as a strategic decision to follow the path with the highest reward probability. Stressed participants rather decided randomly in the first stage to then proceed to the reward-guided choice in the second stage. One possible mechanism is the stress-induced release of noradrenaline promoting explorative behavior by stopping the use of previously collected information in favor of exploring new options (Dayan & Yu, 2006; Dubois et al., 2021). This is further supported by the behavioral results indicating that stressed participants did not show a worse learning performance than the control group per se, neither in the first, nor in the second stage. At the same time, the groups did not differ in their learning rate α in either stage. Another possible mechanism is that participants in the stress group were not able to maintain the relevant information to guide their choices as a consequence of detrimental stress effects on WM. Recent evidence suggests a bimodal distribution of negative stress effects on WM with early effects following elevated noradrenaline levels, and later effects being mediated by non-genomic cortisol effects (Geißler et al., 2023). In line with this finding, our data show particularly random first-stage choice behavior after acute stress in participants with low WM capacity, compared to both stressed participants with high WM capacity and to participants with low and high WM capacity in the control group.

On the neural level, stress reduced signatures of both model-based and model-free prediction errors. With regards to model-based prediction errors, the stress group showed reduced activity in the posterior hippocampus and in the putamen, compared to the control group. The hippocampus is associated with encoding cognitive maps by holding representations between cues, actions, outcomes as well as further characteristics of the environment (Schuck & Niv, 2019; Wikenheiser & Schoenbaum, 2016). Model-free prediction errors on the other hand, were associated with reduced activation in the ilPFC in the stress group in comparison to the control group. Interestingly, for model-based prediction errors, ilPFC activity was increased in the stress group. The ilPFC has been associated with the role of an arbitrator, signaling whether behavioral control should be controlled by the model-based or the model-free RL system (Lee et al., 2014). This mechanism might work by suppressing the model-free system when the reliability of model-based information is assumed to be higher. Together, this might explain the stress-induced increase of ilPFC activity during model-based value computations.

However, it should be noted that both our behavioral and neural results do not show a clear dissociation of a model-based vs. model-free system underlying learning and decision-making in the two-step task. Recent studies cast doubt on the idea of distinct systems contributing differentially to decision-making, arguing that mode-free and model-based algorithms consist of numerous independent computation subcomponents which can be recombined in ways that blur the boundaries between model-based and model-free RL (Collins & Cockburn, 2020). For example, the overall assumption is that the transition function is learned through explicit reasoning, reflected in the process of forming a cognitive map of the task space, the core indicator of model-based behavior. Alternatively, the transition function could also be learned by model-free learning processes which account for discrepancies between expected and observed state transitions (Kurdi et al., 2019). Further, both systems are partially drawing on a common set of computational principles, such as the valuation of action-outcome associations via RPEs that are shared across both model-based and model-free learning systems (Daw et al., 2011; Gläscher et al., 2010). Accordingly, an alternative view could be that model-based learning is an extension of model-free RL rather than being distinct systems. This view is further supported by the lack of a clear dissociation on the neural level. Previous research has identified neural substrates that are activated during both model-based and model-free decision-making processes such as the vmPFC and the striatum (Deserno et al., 2015; Jocham et al., 2011), suggesting that these brain regions support a mix of computational processes associated with both model-based and model-free learning. Additionally, RL algorithms typically assume predefined state and action spaces, but humans and animals often must discover these task spaces (Collins & Cockburn, 2020). Identifying the state space of the current environment likely involves separate networks in the brain, such as the prefrontal cortex and hippocampus. Together, the notion that distinct neural circuits support exclusively model-based or model-free learning should be challenged by a more nuanced view that considers the interaction and overlap of these systems to accurately reflect the complexity of cognitive processes.

3.2 Functionally different roles of dopamine and noradrenaline in exploration and exploitation

The exploration-exploitation tradeoff is a fundamental part of learning and decision-making. It is assumed to arise from dissociable systems on behavioral, computational, and neural levels. While the literature identified dopamine and noradrenaline as key players in signaling choice-relevant aspects when deciding whether to repeat known actions or to explore new options, the exact mechanisms are widely unknown. We sought out to shed light on the specific roles of dopamine and noradrenaline by pharmacologically blocking either system before performing a patch-leaving foraging task. We particularly manipulated the rewards associated with choice options, the degree to which rewards decreased over time, and the opportunity costs it took to switch to a new option to gain insights about the functional roles associated with one or the other system. We hypothesized higher levels of striatal dopamine to induce a focus on reward representations (Pagnoni et al., 2002; Schultz et al., 1997), therefore favoring this aspect to guide choices, resulting in rather exploitative behavior. Prefrontal dopamine on the other hand should be involved in the search for alternative options to gain information, therefore promoting explorative behavior.

Participants in the amisulpride group switched less, especially when (i) the previous reward was high, (ii) the travel time was long, and (iii) the depletion rate was low. Thus, our results provide support for an increased sensitivity towards these choice aspects in the amisulpride group. Given the blocking effects of amisulpride on dopamine, this seems surprising at first glance. However, our results align with a dual-state model of prefrontal dopamine (Seamans et al., 2001). This model proposes that when prefrontal D1 receptors dominate activation, GABAergic inhibition is increased, acting as a gate where only strong inputs can pass to prefrontal circuits. When D2 receptors are primarily active, GABAergic inhibition is decreased, leading to also weak inputs passing to prefrontal circuits (Seamans et al., 2001). Blocking prefrontal D2 receptors may induce a shift towards D1 activation, therefore promoting the processing of strong inputs (Seamans & Yang, 2004). Amisulpride was shown to preferentially block D2/D3 receptors in the PFC, while dopamine levels in the striatum were even elevated after low doses (Bressan et al., 2003; Scatton et al., 1997; Viviani et al., 2013). Thus, our results may reflect a shift towards prefrontal D1 receptor activation within the prefrontal cortex. Combined with an intact striatal dopamine function, this might have promoted the development of robust representations of decision-relevant stimuli and ultimately heightened sensitivity to specific aspects of choice to guide choices.

In contrast to the amisulpride group, the propranolol group showed no significant effects of choice aspects on behavior. Intriguingly, they tended to switch even more after higher rewards, compared to placebo. This indicates that participants in the propranolol group were less sensitive to the decision-relevant information they encountered, consistent with findings proposing a role of noradrenaline in random, but not directed exploration (Dubois et al., 2021; Jahn et al., 2018; Tervo et al., 2014; Warren et al., 2017). However, there are mixed findings regarding the direction of noradrenaline influencing choice randomness. Increased noradrenergic activity was associated with less random exploration (Warren et al., 2017), while other studies linked increased decision noise to higher levels of noradrenaline (Fan et al., 2023; Jepma & Nieuwenhuis, 2011). These heterogenous findings might be attributed to the different activity modes of noradrenaline. Tonic noradrenergic firing was linked to explorative behavior, whereas phasic noradrenaline is believed to promote exploitation (Aston-Jones & Cohen, 2005). Since propranolol is suggested to affect both tonic and phasic signaling of noradrenaline (Lawson et al., 2021), distinguishing between these modes based on our data is not feasible. One potential explanation for the increased stochasticity we found after blocking noradrenaline could be an inhibitory mechanism involving β-adrenergic receptors. Specifically, research indicated that β-adrenergic receptors in rats boosted inhibitory synaptic mechanisms through a noradrenaline-mediated enhancement of GABA efficacy (Waterhouse et al., 1982). By blocking these receptors, we may have disrupted a noradrenaline-related suppression of noise, leading to an increase of decision noise and therefore an increase of random behavior.

Our results imply that it is not one or the other transmitter system that supports exploration or exploitation per se, but that dopamine and noradrenaline each signal specific components of the decision-making process. It can therefore be concluded that dopamine is responsible for processing the decision-relevant information that is used in both exploratory and exploitative behavior, while noradrenaline transmits in particular the continuation or abandonment of the current strategy. In the context of the proposed dual-systems approach of exploration and exploitation, our results indicate that the assumption of two separate systems is questionable. It becomes apparent already in the formalization that exploration and exploitation are essentially based on the same mechanism, that is the accumulation of information in interaction with the environment to assess the quality of the current option to guide future decisions. Behaviorally, exploration and exploitation are indicated by an overall tendency towards switching or staying, respectively and computationally the intercept parameter was introduced to capture constant choice biases. Consequently, it is not assumed that exploration and exploitation arise from fundamentally different processes, but that they are rather two poles on a continuum that are the result of computations of multiple choice-relevant components.

3.3 Cooperative interplay between RL and WM

In the lens of dual-systems accounts the idea to disentangle WM and RL contributions to learning and decision-making is a seminal breakthrough. Given the basic assumption that the RL-immanent learning of value functions requires an active maintenance of stimulus-action-outcome associations, the involvement of WM processes is obvious. We sought out to test the extent to which WM and RL are alternative systems by (i) experimentally manipulating RL and WM demands, (ii) disentangling the signatures indicative of each process in a computational model and (iii), providing insights into the underlying neural systems by decoding EEG signals.
Our behavioral results corroborate previous findings identifying separable contributions of RL and WM to successful learning (Collins, 2018; Collins, Albrecht, et al., 2017; Collins & Frank, 2012). Specifically, we showed that successful learning consisted of an RL process that

incrementally accumulated value information and a WM process specifically present in blocks with low set sizes and prone to delay effects. We further showed that representations learned under high load, therefore processed via RL, are more robust than those learned under low load when WM guided learning, as indicated by an improved recall performance for associations learned under higher set sizes in the stimulus-response retention test (test phase 2). This is in line with previous evidence for this test phase (Collins, 2018). In the same vein, participants were more likely to select stimuli associated with higher rewards during learning in the reward retention test (test phase 1), confirming RL contributions in the learning phase. This value discrimination effect was also enhanced under higher WM loads (replicating Collins, Albrecht, et al., 2017). Importantly, our test phase design enables a measurement of pure RL signatures, as the first test phase specifically tests the incremental accumulation of reward values, and the second test phase takes place ~15 min after the learning phase offset when WM representations are no longer accessible. Taken together, our results support the notion that a higher WM load benefits retention accuracy, aligning with the hypothesis that RL processes, when under the pressure of high WM load, play a pivotal role in ensuring robust retention of learned associations.

This is further confirmed by our model-based EEG-results. Trial-by-trial recordings showed early frontal activity associated with the RL system followed by later parietal signals linked to set size. As demonstrated earlier (Collins & Frank, 2018), these dynamics indicate an early activation of the RL system, followed by the cognitively effortful WM system. Importantly, the neural signature of RL increased with increasing reward history. This effect was even higher with increasing set size, in line with earlier observations of RL overtaking learning when WM contributions decrease (Collins, Albrecht, et al., 2017; Collins & Frank, 2018). Strikingly, our results provide evidence indicating that RL indices during the acquisition phase are predictive of retention performance, highlighting the significance of RL in learning and memory beyond the immediate contributions of WM.

Although stress is known to influence a broad range of cognitive domains (Arnsten, 2009; Luksys & Sandi, 2011; Schwabe et al., 2012), stress had a limited effect on the RL and WM trade-off in our study. In previous studies, stress was found to impair WM performance (Geißler et al., 2023; Qin et al., 2009; Woodcock et al., 2019), while even increasing striatal dopamine activity (Vaessen et al., 2015). We therefore reasoned that stress would lead to a shift from WM-guided processes to RL-drivel learning. In fact, our data only partly support this hypothesis. Acute stress slightly modulated the interaction between RL and WM in the learning phase with a greater performance drop in blocks with high set sizes, compared to the control group. However, stress did not have an impact on the WM indices alone. Moreover, we found a hint for a stress × recall accuracy × set size effect in the stimulus-response retention test, suggesting acute stress might impact the retention of learned stimulus-response associations differently based on the WM load during learning. However, follow-up analyses could not confirm a robust stress effect in the second test phase. We further did not find significant stress effects on the performance in the reward retention test, but as WM is not the primary component in this phase, this test phase was not the center of our hypothesis. The lack of stress effects cannot be attributed to our manipulation per se, as subjective mood ratings, as well as blood pressure and cortisol measures clearly indicated a stress reaction after being exposed to the stress manipulation. An attenuation of stress effects on task performance might be explained by the timeline of our experiment. The learning phase began 25min after stressor onset. In this time interval, prefrontal noradrenaline levels are expected to have returned to baseline (peaking around 10 min after stressor offset; Geißler et al., 2023), while central cortisol levels peak at 25-30min after stressor onset (Schwabe & Schächinger, 2018). It is suggested that both noradrenaline *and* cortisol levels need to be increased to find detrimental stress effects on performance (Barsegyan et al., 2010; Elzinga & Roelofs, 2005; Roozendaal et al., 2006). Another possibility is that individual WM capacity influenced how stress affected task performance, as a high WM capacity prevented detrimental stress effects in earlier studies (Quaedflieg et al., 2019).

Together, we found that a higher WM load led to a shift towards RL-guided learning that was less time-sensitive and therefore more robust. This was further backed by our neural results indicating that RL indices during the acquisition phase were predictive of retention performance, highlighting the cooperative interplay between WM and RL when guiding decision-making. The integration of WM into models of RL is an important step towards the consideration of fundamental cognitive processes in studying learning and decision-making. However, it is questionable whether a dichotomization approach is the optimal solution, which again neglects important processes such as attention or motor learning.

3.4 Re-evaluating the notion of dual-systems approaches in RL

Taken together, our studies jointly shed light on the nuanced interplay within RL processes, each emphasizing key aspects underlying RL contributions to decision-making. In the study of model-based versus model-free learning, we found that individuals utilize both strategies, with a tendency towards model-free learning. Stress, rather than specifically impairing model-based learning, reduced signatures of both model-based and model-free RL, paralleled by an overall tendency towards increased choice randomness. In Study 2, we showed that dopamine and noradrenaline play functionally distinct roles in the exploration-exploitation tradeoff. Dopamine was shown to be crucial for processing information relevant to choices, while noradrenaline appeared to affect decision-making more generally by regulating when to disengage from the current information paths to randomly explore new options. Finally, we examined the interplay between WM and RL, uncovering a synergistic interaction where high WM demands enhanced the retention accuracy of learned associations. This suggests that WM and RL processes, rather than acting as alternative systems, collaboratively contribute to learning efficiency and decision-making robustness.

Although the approaches examined here are each represented in their own niches of the literature, they exhibit considerable overlaps in their behavioral, computational, and neural mechanisms (Figures 4, 7, and 9). Notably, all tasks share the same core of learning stimulus-action pairs associated with rewards to guide choice behavior. Therefore, they are all grounded in the same computational framework wherein the basic RL module defines learning as capturing a value function for available stimulus-action-associations (Gershman & Uchida, 2019). RPEs enable value updates while interacting with the environment, adjusted by learning rates that determine the extent to which new information is used to update the value representations (Sutton & Barto, 1998). Additionally, inverse temperature parameters play a pivotal role, indicating the extent to which the acquired knowledge guides decision-making processes. At the same time, all three approaches propose similar neural processes and brain regions guiding learning and decision making (Figures 4, 7, and 9) with the involvement of dopaminergic RPE computations in the striatum, ventral/dorsal prefrontal activity, and parietal regions. In conclusion, all three approaches explored in this thesis, while initially situated within distinct areas of research, reveal significant overlaps in their behavioral, computational, and neural underpinnings.

Importantly, the integration of the results from the three studies provides insights beyond the isolated view within the respective approaches. Both Study 1 and Study 2 revealed a behavioral signature where participants disengaged the path of accumulating information to guide behavior, but instead performed value-independent, random choices. In the first study, this behavior was particularly present in the stress group, in the second study participants in the propranolol group showed an increase of randomness. In light of the well-established release of noradrenaline as part of the acute stress response (Joëls & Baram, 2009), it is tempting to speculate that noradrenaline mirrored a reset signal leading to the disengagement from previously learned information in both studies. We further showed that random choice behavior in Study 1 was specifically pronounced in stressed participants with a low WM capacity, suggesting that the underlying mechanism was not reflected in a reset-signal, but rather a consequence of detrimental stress effects on WM, therefore explained by an inability of maintaining choice-relevant information. Taking into account that the key characteristic of model-based RL is the creation and application of an environmental model, the importance of WM contributions seems specifically apparent for model-based RL. However, as we found a general bias towards model-free processes in our study, we cannot make this differentiation. Previous studies have already shown this relationship (Otto et al., 2013; Sharp et al., 2016), so that a further investigation of WM involvement in model-based versus model-free learning processes might be particularly interesting. Together, we show the importance of noradrenaline in signaling a disengagement from the current strategy, we highlight the role of dopamine in processing choice-relevant aspects, and we present striking evidence for a cooperative interplay between WM and RL.

However, although we were able to confirm the signatures of the assumed systems in each of our studies, the substantial parallels between the frameworks also raise doubts about their general validity. This is particularly evident in the concepts of model-based vs. model-free RL and exploration vs. exploitation. Although both approaches are based on different tasks and models, they significantly overlap. For example, when creating an environmental model which is the core process of model-based RL, it is essential to obtain information by trying out unknown paths, a process commonly referred to as directed exploration (Gershman & Tzovaras, 2018; Wiehler et al., 2021). At the same time, model-free RL parallels an exploitative strategy, where agents rely on learned values or policies derived directly from the outcomes of their actions. This mirrors exploitation behavior, where actions are chosen that are associated with the maximum known reward. Despite the obvious parallels, the literature currently mostly does not integrate these approaches. The dedicated assessment of WM processes is a step in the direction of a more integrative perspective, but nevertheless this is also an isolated view on a priori defined processes. While such simplifications might be appropriate in fundamental research, with the increasing complexity of a task, it will in turn be subject to the criticism that important influences of other systems are being ignored, such as attention, episodic memory, or motor learning.

Finally, the results presented in this work also give rise on critical aspects associated with dual-systems approaches in RL. First, the theoretical foundation of model-based vs. model-free RL as well as the exploration-exploitation formalization does not clearly map onto separable systems, but rather draws on (partly) shared mechanisms. Consequently, it would be more accurate to speak of model-based RL as an extension of model-free RL, and of two poles on a continuum in the case of exploration and exploitation. Second, this argument is supported by empirical results including the ones presented here that show that the respective systems do not occur in their pure form, neither at the behavioral nor at the neural level. Thirdly, the models can only depict the space limited by their own formalization, which represents an under-complex approximation to high dimensional decision problems.

3.5 Conclusion and future directions

Together, our studies challenge the traditional dichotomy of learning systems proposed in earlier RL theories. Instead, they highlight a more integrated and dynamic interplay of cognitive processes, suggesting that learning and decision-making behaviors emerge from the complex interactions between multiple systems rather than being driven by an isolated selection of dichotomized systems. At the same time, we show considerable overlaps in the behavioral, computational, and neural mechanisms between the approaches, although the individual frameworks are mainly considered independent from each other in the literature. Recent advances criticize that predefined computational models can only reflect the process that is based on the underlying theory (Eckstein et al., 2021), therefore nurturing the existence of separate niches for similar objectives. Moreover, it is proposed that such "handcrafted" models come with the risk of an incorrect specification of the model, meaning for example that mechanisms critical for the actual data generation process can easily be overlooked (Nassar & Frank, 2016). An alternative framework is offered by deep learning approaches where models are defined by raw data input, therefore allowing the agent to adapt and optimize choice strategies in complex and dynamic environments (Cross et al., 2021). Although dichotomized systems have a long tradition in cognitive science and have provided groundbreaking insights, future studies should take a more integrative view on the complex interplay of multiple systems underlying learning and decision-making.

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

Anstrom, K. K., & Woodward, D. J. (2005). Restraint Increases Dopaminergic Burst Firing in Awake Rats. *Neuropsychopharmacology*, *30*(10), 1832–1840. https://doi.org/10.1038/sj.npp.1300730

Arnsten, A. F. T. (2009). Stress signalling pathways that impair prefrontal cortex structure and function. *Nature Reviews Neuroscience*, *10*(6), 410–422. https://doi.org/10.1038/nrn2648

Aston-Jones, G., & Cohen, J. D. (2005). AN INTEGRATIVE THEORY OF LOCUS COERULEUS-NOREPINEPHRINE FUNCTION: Adaptive Gain and Optimal Performance. *Annual Review of Neuroscience*, *28*(1), 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709

Averbeck, B. B. (2015). Theory of Choice in Bandit, Information Sampling and Foraging Tasks. *PLOS Computational Biology*, *11*(3), e1004164. https://doi.org/10.1371/journal.pcbi.1004164

Averbeck, B., & O'Doherty, J. P. (2022). Reinforcement-learning in fronto-striatal circuits. *Neuropsychopharmacology*, *47*(1), 147–162.

Baddeley, A. (1992). Working Memory. *Science*, *255*(5044), 556–559. https://doi.org/10.1126/science.1736359

Badre, D., Doll, B. B., Long, N. M., & Frank, M. J. (2012). Rostrolateral Prefrontal Cortex and Individual Differences in Uncertainty-Driven Exploration. *Neuron*, *73*(3), 595–607. https://doi.org/10.1016/j.neuron.2011.12.025

Balleine, B. W., & O'Doherty, J. P. (2010). Human and Rodent Homologies in Action Control: Corticostriatal Determinants of Goal-Directed and Habitual Action. *Neuropsychopharmacology*, *35*(1), 48–69. https://doi.org/10.1038/npp.2009.131

Balleine, B. W., & Ostlund, S. B. (2007). Still at the choice-point: Action selection and initiation in instrumental conditioning. *Annals of the New York Academy of Sciences*, *1104*(1), 147–171.

Barsegyan, A., Mackenzie, S. M., Kurose, B. D., McGaugh, J. L., & Roozendaal, B. (2010). Glucocorticoids in the prefrontal cortex enhance memory consolidation and impair working memory by a common neural mechanism. *Proceedings of the National Academy of Sciences*, *107*(38), 16655–16660. https://doi.org/10.1073/pnas.1011975107

Barto, A. G. (2003). Reinforcement Learning in Motor Control. In *Handbook of brain theory and neural networks* (Bd. 2, S. 968–972).

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal. *Neuron*, *47*(1), 129–141. https://doi.org/10.1016/j.neuron.2005.05.020

Beharelle, A. R., Polania, R., Hare, T. A., & Ruff, C. C. (2015). Transcranial Stimulation over Frontopolar Cortex Elucidates the Choice Attributes and Neural Mechanisms Used to Resolve Exploration-Exploitation Trade-Offs. *Journal of Neuroscience*, *35*(43), 14544–14556. https://doi.org/10.1523/JNEUROSCI.2322-15.2015

Berryhill, M. E., Chein, J., & Olson, I. R. (2011). At the intersection of attention and memory: The mechanistic role of the posterior parietal lobe in working memory. *Neuropsychologia*, *49*(5), 1306–1315. https://doi.org/10.1016/j.neuropsychologia.2011.02.033

Birnbaum, S. G., Yuan, P. X., Wang, M., Vijayraghavan, S., Bloom, A. K., Davis, D. J., Gobeske, K. T., Sweatt, J. D., Manji, H. K., & Arnsten, A. F. T. (2004). Protein Kinase C Overactivity Impairs Prefrontal Cortical Regulation of Working Memory. *Science*, *306*(5697), 882–884. https://doi.org/10.1126/science.1100021

Blanchard, T. C., & Gershman, S. J. (2018). Pure correlates of exploration and exploitation in the human brain. *Cognitive, Affective, & Behavioral Neuroscience*, *18*(1), 117–126. https://doi.org/10.3758/s13415-017-0556-2

Boorman, E. D., Behrens, T. E. J., Woolrich, M. W., & Rushworth, M. F. S. (2009). How Green Is the Grass on the Other Side? Frontopolar Cortex and the Evidence in Favor of Alternative Courses of Action. *Neuron*, *62*(5), 733–743. https://doi.org/10.1016/j.neuron.2009.05.014

Bouton, M. E., Maren, S., & McNally, G. P. (2021). Behavioral and neurobiological mechanisms of pavlovian and instrumental extinction learning. *Physiological Reviews*, *101*(2), 611–681. https://doi.org/10.1152/physrev.00016.2020

Bressan, R. A., Erlandsson, K., Jones, H. M., Mulligan, R., Flanagan, R. J., Ell, P. J., & Pilowsky, L. S. (2003). Is regionally selective D2/D3 dopamine occupancy sufficient for atypical antipsychotic effect? An in vivo quantitative [123I] epidepride SPET study of amisulpride-treated patients. *American Journal of Psychiatry*, *160*(8), 1413–1420. https://doi.org/10.1176/appi.ajp.160.8.1413

Camille, N., Tsuchida, A., & Fellows, L. K. (2011). Double Dissociation of Stimulus-Value and Action-Value Learning in Humans with Orbitofrontal or Anterior Cingulate Cortex Damage. *The Journal of Neuroscience*, *31*(42), 15048–15052. https://doi.org/10.1523/JNEUROSCI.3164-11.2011

Cerqueira, J. J., Mailliet, F., Almeida, O. F. X., Jay, T. M., & Sousa, N. (2007). The Prefrontal Cortex as a Key Target of the Maladaptive Response to Stress. *The Journal of Neuroscience*, *27*(11), 2781–2787. https://doi.org/10.1523/JNEUROSCI.4372-06.2007

Chakroun, K., Mathar, D., Wiehler, A., Ganzer, F., & Peters, J. (2020). Dopaminergic modulation of the exploration/exploitation trade-off in human decision-making. *eLife*, *9*. https://doi.org/10.7554/eLife.51260

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology*, *9*(2), 129–136. https://doi.org/10.1016/0040-5809(76)90040-X

Chatham, C. H., Frank, M. J., & Badre, D. (2014). Corticostriatal Output Gating during Selection from Working Memory. *Neuron*, *81*(4), 930–942. https://doi.org/10.1016/j.neuron.2014.01.002

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), 933–942. https://doi.org/10.1098/rstb.2007.2098

Cohen, J. D., Perlstein, W. M., Braver, T. S., Nystrom, L. E., Noll, D. C., Jonides, J., & Smith, E. E. (1997). Temporal dynamics of brain activation during a working memory task. *Nature*, *386*(6625), 604–608.

Collins, A. G. E. (2018). The Tortoise and the Hare: Interactions between Reinforcement Learning and Working Memory. *Journal of Cognitive Neuroscience*, *30*(10), 1422–1432. https://doi.org/10.1162/jocn_a_01238

Collins, A. G. E., Albrecht, M. A., Waltz, J. A., Gold, J. M., & Frank, M. J. (2017). Interactions Among Working Memory, Reinforcement Learning, and Effort in Value-Based Choice: A New Paradigm and Selective Deficits in Schizophrenia. *Biological Psychiatry*, *82*(6), 431–439. https://doi.org/10.1016/j.biopsych.2017.05.017

Collins, A. G. E., Brown, J. K., Gold, J. M., Waltz, J. A., & Frank, M. J. (2014). Working Memory Contributions to Reinforcement Learning Impairments in Schizophrenia. *The Journal of Neuroscience*, *34*(41), 13747–13756. https://doi.org/10.1523/JNEUROSCI.0989-14.2014

Collins, A. G. E., Ciullo, B., Frank, M. J., & Badre, D. (2017). Working Memory Load Strengthens Reward Prediction Errors. *The Journal of Neuroscience*, *37*(16), 4332–4342. https://doi.org/10.1523/JNEUROSCI.2700-16.2017

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

Collins, A. G. E., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis: Working memory in reinforcement learning. *European Journal of Neuroscience*, *35*(7), 1024–1035. https://doi.org/10.1111/j.1460-9568.2011.07980.x

Collins, A. G. E., & Frank, M. J. (2018). Within- and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. *Proceedings of the National Academy of Sciences*, *115*(10), 2502–2507. https://doi.org/10.1073/pnas.1720963115

Constantino, S. M., & Daw, N. D. (2015). Learning the opportunity cost of time in a patch-foraging task. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(4), 837–853. https://doi.org/10.3758/s13415-015-0350-y

Cools, R., & D'Esposito, M. (2011). Inverted-U–Shaped Dopamine Actions on Human Working Memory and Cognitive Control. *Biological Psychiatry*, *69*(12), e113–e125. https://doi.org/10.1016/j.biopsych.2011.03.028

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114. https://doi.org/10.1017/S0140525X01003922

Cowan, N. (2008). Chapter 20 What are the differences between long-term, short-term, and working memory? In *Progress in Brain Research* (Bd. 169, S. 323–338). Elsevier. https://doi.org/10.1016/S0079-6123(07)00020-9

Cross, L., Cockburn, J., Yue, Y., & O'Doherty, J. P. (2021). Using deep reinforcement learning to reveal how the brain encodes abstract state-space representations in high-dimensional environments. *Neuron*, *109*(4), 724-738.e7. https://doi.org/10.1016/j.neuron.2020.11.021

Daw, N. D., & Doya, K. (2006). The computational neurobiology of learning and reward. *Current Opinion in Neurobiology*, *16*(2), 199–204. https://doi.org/10.1016/j.conb.2006.03.006

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. (2005a). Actions, Policies, Values, and the Basal Ganglia. *Recent Breakthroughs in Basal Ganglia Research*, *10*(9), 1214–1221.

Daw, N. D., Niv, Y., & Dayan, P. (2005b). 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., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879. https://doi.org/10.1038/nature04766

Dayan, P., & Balleine, B. W. (2002). Reward, Motivation, and Reinforcement Learning. *Neuron*, *36*(2), 285–298. https://doi.org/10.1016/S0896-6273(02)00963-7

Dayan, P., & Yu, A. J. (2006). Phasic norepinephrine: A neural interrupt signal for unexpected events. *Network: Computation in Neural Systems*, *17*(4), 335–350. https://doi.org/10.1080/09548980601004024

De Kloet, E. R., Joëls, M., & Holsboer, F. (2005). Stress and the brain: From adaptation to disease. *Nature Reviews Neuroscience*, *6*(6), 463–475. https://doi.org/10.1038/nrn1683

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, *16*(1), 105–110. https://doi.org/10.1038/nn.3279

Deserno, L., Huys, Q. J. M., Boehme, R., Buchert, R., Heinze, H.-J., Grace, A. A., Dolan, R. J., Heinz, A., & Schlagenhauf, F. (2015). Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. *Proceedings of the National Academy of Sciences*, *112*(5), 1595–1600. https://doi.org/10.1073/pnas.1417219112

D'Esposito, M., & Postle, B. R. (2015). The Cognitive Neuroscience of Working Memory. *Annual Review of Psychology*, *66*(1), 115–142. https://doi.org/10.1146/annurev-psych-010814-015031

Dolan, R. J., & Dayan, P. (2013). Goals and Habits in the Brain. *Neuron*, *80*(2), 312–325. https://doi.org/10.1016/j.neuron.2013.09.007

Doll, B. B., Bath, K. G., Daw, N. D., & Frank, M. J. (2016). Variability in Dopamine Genes Dissociates Model-Based and Model-Free Reinforcement Learning. *Journal of Neuroscience*, *36*(4), 1211–1222. https://doi.org/10.1523/JNEUROSCI.1901-15.2016

Doody, M., Van Swieten, M. M. H., & Manohar, S. G. (2022). Model-based learning retrospectively updates model-free values. *Scientific Reports*, *12*(1), 2358. https://doi.org/10.1038/s41598-022-05567-3

Drummond, N., & Niv, Y. (2020). Model-based decision making and model-free learning. *Current Biology*, *30*(15), R860–R865. https://doi.org/10.1016/j.cub.2020.06.051

Dubois, M., Habicht, J., Michely, J., Moran, R., Dolan, R. J., & Hauser, T. U. (2021). Human complex exploration strategies are enriched by noradrenaline-modulated heuristics. *eLife*, *10*. https://doi.org/10.7554/eLife.59907

Durstewitz, D., Seamans, J. K., & Sejnowski, T. J. (2000). Neurocomputational models of working memory. *Nature Neuroscience*, *3*(S11), 1184–1191. https://doi.org/10.1038/81460

Eckstein, M. K., Wilbrecht, L., & Collins, A. G. (2021). What do reinforcement learning models measure? Interpreting model parameters in cognition and neuroscience. *Current Opinion in Behavioral Sciences*, *41*, 128–137. https://doi.org/10.1016/j.cobeha.2021.06.004

Elzinga, B. M., & Roelofs, K. (2005). Cortisol-Induced Impairments of Working Memory Require Acute Sympathetic Activation. *Behavioral Neuroscience*, *119*(1), 98–103. https://doi.org/10.1037/0735-7044.119.1.98

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Fan, H., Burke, T., Sambrano, D. C., Dial, E., Phelps, E. A., & Gershman, S. J. (2023). Pupil Size Encodes Uncertainty during Exploration. *Journal of Cognitive Neuroscience*, *35*(9), 1508–1520. https://doi.org/10.1162/jocn_a_02025

Frank, M. J., Doll, B. B., Oas-Terpstra, J., & Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience*, *12*(8), 1062–1068. https://doi.org/10.1038/nn.2342

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism. *Science*, *306*(5703), 1940–1943. https://doi.org/10.1126/science.1102941

Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology*, *61*(2), 331–349. https://doi.org/10.1152/jn.1989.61.2.331

Fuster, J. M. (1973). Unit activity in prefrontal cortex during delayed-response performance: Neuronal correlates of transient memory. *Journal of Neurophysiology*, *36*(1), 61–78. https://doi.org/10.1152/jn.1973.36.1.61

Garvert, M. M., Dolan, R. J., & Behrens, T. E. (2017). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *eLife*, *6*. https://doi.org/10.7554/eLife.17086

Geißler, C. F., Friehs, M. A., Frings, C., & Domes, G. (2023). Time-dependent effects of acute stress on working memory performance: A systematic review and hypothesis. *Psychoneuroendocrinology*, *148*, 105998. https://doi.org/10.1016/j.psyneuen.2022.105998

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42. https://doi.org/10.1016/j.cognition.2017.12.014

Gershman, S. J., & Daw, N. D. (2017). Reinforcement Learning and Episodic Memory in Humans and Animals: An Integrative Framework. *Annual Review of Psychology*, *68*(1), 101–128. https://doi.org/10.1146/annurev-psych-122414-033625

Gershman, S. J., & Tzovaras, B. G. (2018). Dopaminergic genes are associated with both directed and random exploration. *Neuropsychologia*, *120*, 97–104. https://doi.org/10.1016/j.neuropsychologia.2018.10.009

Gershman, S. J., & Uchida, N. (2019). Believing in dopamine. *Nature Reviews Neuroscience*, *20*(11), 703–714. https://doi.org/10.1038/s41583-019-0220-7

Gigerenzer, G. (2010). Personal Reflections on Theory and Psychology. *Theory & Psychology*, *20*(6), 733–743. https://doi.org/10.1177/0959354310378184

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. *Neuron*, *66*(4), 585–595. https://doi.org/10.1016/j.neuron.2010.04.016

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, *108*(supplement_3), 15647–15654. https://doi.org/10.1073/pnas.1014269108

Goldman-Rakic, P. S. (1995). Cellular basis of working memory. *Neuron*, *14*(3), 477–485. https://doi.org/10.1016/0896-6273(95)90304-6

Groman, S. M., Keistler, C., Keip, A. J., Hammarlund, E., DiLeone, R. J., Pittenger, C., Lee, D., & Taylor, J. R. (2019). Orbitofrontal Circuits Control Multiple Reinforcement-Learning Processes. *Neuron*. https://doi.org/10.1016/j.neuron.2019.05.042

Gruber, A. J., Dayan, P., Gutkin, B. S., & Solla, S. A. (2006). Dopamine modulation in the basal ganglia locks the gate to working memory. *Journal of Computational Neuroscience*, *20*(2), 153–166. https://doi.org/10.1007/s10827-005-5705-x

Halsey, L. G., & Butler, P. J. (2006). Optimal diving behaviour and respiratory gas exchange in birds. *Respiratory Physiology & Neurobiology*, *154*(1–2), 268–283. https://doi.org/10.1016/j.resp.2006.01.012

Hauser, T. U., Iannaccone, R., Ball, J., Mathys, C., Brandeis, D., Walitza, S., & Brem, S. (2014). Role of the Medial Prefrontal Cortex in Impaired Decision Making in Juvenile Attention-Deficit/Hyperactivity Disorder. *JAMA Psychiatry*, *71*(10), 1165. https://doi.org/10.1001/jamapsychiatry.2014.1093

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2006). Banishing the homunculus: Making working memory work. *Neuroscience*, *139*(1), 105–118. https://doi.org/10.1016/j.neuroscience.2005.04.067

Hermans, E. J., Henckens, M. J. A. G., Joëls, M., & Fernández, G. (2014). Dynamic adaptation of large-scale brain networks in response to acute stressors. *Trends in Neurosciences*, *37*(6), 304–314. https://doi.org/10.1016/j.tins.2014.03.006

Hogeveen, J., Mullins, T. S., Romero, J. D., Eversole, E., Rogge-Obando, K., Mayer, A. R., & Costa, V. D. (2022). The neurocomputational bases of explore-exploit decision-making. *Neuron*, *110*(11), 1869-1879.e5. https://doi.org/10.1016/j.neuron.2022.03.014

Holmes, A., & Wellman, C. L. (2009). Stress-induced prefrontal reorganization and executive dysfunction in rodents. *Neuroscience & Biobehavioral Reviews*, *33*(6), 773–783. https://doi.org/10.1016/j.neubiorev.2008.11.005

Jahn, C. I., Gilardeau, S., Varazzani, C., Blain, B., Sallet, J., Walton, M. E., & Bouret, S. (2018). Dual contributions of noradrenaline to behavioural flexibility and motivation. *Psychopharmacology*, *235*(9), 2687–2702. https://doi.org/10.1007/s00213-018-4963-z

Jepma, M., & Nieuwenhuis, S. (2011). Pupil Diameter Predicts Changes in the Exploration–Exploitation Trade-off: Evidence for the Adaptive Gain Theory. *Journal of Cognitive Neuroscience*, *23*(7), 1587–1596. https://doi.org/10.1162/jocn.2010.21548

Jocham, G., Klein, T. A., & Ullsperger, M. (2011). Dopamine-Mediated Reinforcement Learning Signals in the Striatum and Ventromedial Prefrontal Cortex Underlie Value-Based Choices. *Journal of Neuroscience*, *31*(5), 1606–1613. https://doi.org/10.1523/JNEUROSCI.3904-10.2011

Joëls, M., & Baram, T. Z. (2009). The neuro-symphony of stress. *Nature Reviews Neuroscience*, *10*(6), 459–466. https://doi.org/10.1038/nrn2632

Kakade, S., & Dayan, P. (2002). Dopamine: Generalization and bonuses. *Neural Networks*, *15*(4–6), 549–559. https://doi.org/10.1016/S0893-6080(02)00048-5

Kane, G. A., Vazey, E. M., Wilson, R. C., Shenhav, A., Daw, N. D., Aston-Jones, G., & Cohen, J. D. (2017). Increased locus coeruleus tonic activity causes disengagement from a patch-foraging task. *Cognitive, Affective, & Behavioral Neuroscience*, *17*(6), 1073–1083. https://doi.org/10.3758/s13415-017-0531-y

Kim, H., Shimojo, S., & O'Doherty, J. P. (2011). Overlapping Responses for the Expectation of Juice and Money Rewards in Human Ventromedial Prefrontal Cortex. *Cerebral Cortex*, *21*(4), 769–776. https://doi.org/10.1093/cercor/bhq145

Kim, J. J., & Diamond, D. M. (2002). The stressed hippocampus, synaptic plasticity and lost memories. *Nature Reviews Neuroscience*, *3*(6), 453–462. https://doi.org/10.1038/nrn849

Kirschbaum, C., Pirke, K. M., & Hellhammer, D. H. (1993). 'Trier Social Stress Test'–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, *28*(1–2), 76–81.

Knutson, B., & Cooper, J. C. (2005). Functional magnetic resonance imaging of reward prediction: *Current Opinion in Neurology*, *18*(4), 411–417. https://doi.org/10.1097/01.wco.0000173463.24758.f6

Kolling, N., Behrens, T. E. J., Mars, R. B., & Rushworth, M. F. S. (2012). Neural Mechanisms of Foraging. *Science*, *336*(6077), 95–98. https://doi.org/10.1126/science.1216930

Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review*, *118*(1), 97–109. https://doi.org/10.1037/a0020762

Kurdi, B., Gershman, S. J., & Banaji, M. R. (2019). Model-free and model-based learning processes in the updating of explicit and implicit evaluations. *Proceedings of the National Academy of Sciences*, *116*(13), 6035–6044. https://doi.org/10.1073/pnas.1820238116

Laureiro-Martínez, D., Brusoni, S., Canessa, N., & Zollo, M. (2015). Understanding the exploration–exploitation dilemma: An fMRI study of attention control and decision-making performance. *Strat. Manag. J.*, *36*, 319–338.

Lawson, R. P., Bisby, J., Nord, C. L., Burgess, N., & Rees, G. (2021). The Computational, Pharmacological, and Physiological Determinants of Sensory Learning under Uncertainty. *Current Biology*, *31*(1), 163-172.e4. https://doi.org/10.1016/j.cub.2020.10.043

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

Lee, T. G., & D'Esposito, M. (2012). The Dynamic Nature of Top-Down Signals Originating from Prefrontal Cortex: A Combined fMRI–TMS Study. *The Journal of Neuroscience*, *32*(44), 15458–15466. https://doi.org/10.1523/JNEUROSCI.0627-12.2012

Lenow, J. K., Constantino, S. M., Daw, N. D., & Phelps, E. A. (2017). Chronic and Acute Stress Promote Overexploitation in Serial Decision Making. *The Journal of Neuroscience*, *37*(23), 5681–5689. https://doi.org/10.1523/JNEUROSCI.3618-16.2017

Lorenz, R. C., Gleich, T., Beck, A., Pöhland, L., Raufelder, D., Sommer, W., Rapp, M. A., Kühn, S., & Gallinat, J. (2014). Reward anticipation in the adolescent and aging brain. *Human Brain Mapping*, *35*(10), 5153–5165. https://doi.org/10.1002/hbm.22540

Luksys, G., & Sandi, C. (2011). Neural mechanisms and computations underlying stress effects on learning and memory. *Current Opinion in Neurobiology*, *21*(3), 502–508. https://doi.org/10.1016/j.conb.2011.03.003

Marzuki, A. A., Tomić, I., Ip, S. H. Y., Gottwald, J., Kanen, J. W., Kaser, M., Sule, A., Conway-Morris, A., Sahakian, B. J., & Robbins, T. W. (2021). Association of Environmental Uncertainty With Altered Decision-making and Learning Mechanisms in Youths With Obsessive-Compulsive Disorder. *JAMA Network Open*, *4*(11), e2136195. https://doi.org/10.1001/jamanetworkopen.2021.36195

McDougle, S. D., & Collins, A. G. E. (2021). Modeling the influence of working memory, reinforcement, and action uncertainty on reaction time and choice during instrumental learning. *Psychonomic Bulletin & Review*, *28*(1), 20–39. https://doi.org/10.3758/s13423-020-01774-z

McEWEN, B. S. (1998). Stress, Adaptation, and Disease: Allostasis and Allostatic Load. *Annals of the New York Academy of Sciences*, *840*(1), 33–44. https://doi.org/10.1111/j.1749-6632.1998.tb09546.x

McEwen, B. S. (2007). Physiology and Neurobiology of Stress and Adaptation: Central Role of the Brain. *Physiological Reviews*, *87*(3), 873–904. https://doi.org/10.1152/physrev.00041.2006

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), 191–215. https://doi.org/10.1037/dec0000033

Miller, E. K., Lundqvist, M., & Bastos, A. M. (2018). Working Memory 2.0. *Neuron*, *100*(2), 463–475. https://doi.org/10.1016/j.neuron.2018.09.023

Miller, K. J., Botvinick, M. M., & Brody, C. D. (2017). Dorsal hippocampus contributes to model-based planning. *Nature Neuroscience*, *20*(9), 1269–1276. https://doi.org/10.1038/nn.4613

Mobbs, D., Trimmer, P. C., Blumstein, D. T., & Dayan, P. (2018). Foraging for foundations in decision neuroscience: Insights from ethology. *Nature Reviews Neuroscience*, *19*(7), 419–427. https://doi.org/10.1038/s41583-018-0010-7

Möhring, L., & Gläscher, J. (2023). Prediction errors drive dynamic changes in neural patterns that guide behavior. *Cell Reports*, *42*(8), 112931. https://doi.org/10.1016/j.celrep.2023.112931

Murray, J. D., Jaramillo, J., & Wang, X.-J. (2017). Working Memory and Decision-Making in a Frontoparietal Circuit Model. *The Journal of Neuroscience*, *37*(50), 12167–12186. https://doi.org/10.1523/JNEUROSCI.0343-17.2017

Nassar, M. R., & Frank, M. J. (2016). Taming the beast: Extracting generalizable knowledge from computational models of cognition. *Current Opinion in Behavioral Sciences*, *11*, 49–54. https://doi.org/10.1016/j.cobeha.2016.04.003

Oberauer, K., Farrell, S., Jarrold, C., & Lewandowsky, S. (2016). What limits working memory capacity? *Psychological Bulletin*, *142*(7), 758–799. https://doi.org/10.1037/bul0000046

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning. *Science*, *304*(5669), 452–454. https://doi.org/10.1126/science.1094285

O'Doherty, J. P. (2011). Contributions of the ventromedial prefrontal cortex to goal-directed action selection: vmPFC in goal-directed action. *Annals of the New York Academy of Sciences*, *1239*(1), 118–129. https://doi.org/10.1111/j.1749-6632.2011.06290.x

Ossewaarde, L., Qin, S., Van Marle, H. J. F., Van Wingen, G. A., Fernández, G., & Hermans, E. J. (2011). Stress-induced reduction in reward-related prefrontal cortex function. *NeuroImage*, *55*(1), 345–352. https://doi.org/10.1016/j.neuroimage.2010.11.068

Ostlund, S. B., & Balleine, B. W. (2007). Orbitofrontal Cortex Mediates Outcome Encoding in Pavlovian But Not Instrumental Conditioning. *The Journal of Neuroscience*, *27*(18), 4819–4825. https://doi.org/10.1523/JNEUROSCI.5443-06.2007

Ott, T., & Nieder, A. (2019). Dopamine and Cognitive Control in Prefrontal Cortex. *Trends in Cognitive Sciences*, *23*(3), 213–234. https://doi.org/10.1016/j.tics.2018.12.006

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*, *110*(52), 20941–20946. https://doi.org/10.1073/pnas.1312011110

Padoa-Schioppa, C., & Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature*, *441*(7090), 223–226. https://doi.org/10.1038/nature04676

Pagnoni, G., Zink, C. F., Montague, P. R., & Berns, G. S. (2002). Activity in human ventral striatum locked to errors of reward prediction. *Nature Neuroscience*, *5*(2), 97–98. https://doi.org/10.1038/nn802

Park, H., Lee, D., & Chey, J. (2017). Stress enhances model-free reinforcement learning only after negative outcome. *PLOS ONE*, *12*(7), e0180588. https://doi.org/10.1371/journal.pone.0180588

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, *442*(7106), 1042–1045. https://doi.org/10.1038/nature05051

Pfeiffer, B. E., & Foster, D. J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*, *497*(7447), 74–79. https://doi.org/10.1038/nature12112

Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, *400*(6741), 233–238. https://doi.org/10.1038/22268

Porcelli, A. J., & Delgado, M. R. (2017). Stress and decision making: Effects on valuation, learning, and risk-taking. *Current Opinion in Behavioral Sciences*, *14*, 33–39. https://doi.org/10.1016/j.cobeha.2016.11.015

Qin, S., Hermans, E. J., Van Marle, H. J. F., Luo, J., & Fernández, G. (2009). Acute Psychological Stress Reduces Working Memory-Related Activity in the Dorsolateral Prefrontal Cortex. *Biological Psychiatry*, *66*(1), 25–32. https://doi.org/10.1016/j.biopsych.2009.03.006

Quaedflieg, C. W. E. M., Stoffregen, H., Sebalo, I., & Smeets, T. (2019). Stress-induced impairment in goal-directed instrumental behaviour is moderated by baseline working memory. *Neurobiology of Learning and Memory*, *158*, 42–49. https://doi.org/10.1016/j.nlm.2019.01.010

Quilodran, R., Rothé, M., & Procyk, E. (2008). Behavioral Shifts and Action Valuation in the Anterior Cingulate Cortex. *Neuron*, *57*(2), 314–325. https://doi.org/10.1016/j.neuron.2007.11.031

Rac-Lubashevsky, R., & Frank, M. J. (2021). Analogous computations in working memory input, output and motor gating: Electrophysiological and computational modeling evidence. *PLOS Computational Biology*, *17*(6), e1008971. https://doi.org/10.1371/journal.pcbi.1008971

Raio, C. M., Hartley, C. A., Orederu, T. A., Li, J., & Phelps, E. A. (2017). Stress attenuates the flexible updating of aversive value. *Proceedings of the National Academy of Sciences*, *114*(42), 11241–11246. https://doi.org/10.1073/pnas.1702565114

Rajkowski, J., Kubiak, P., Ivanova, S., & Aston-Jones, G. (1997). State-Related Activity, Reactivity of Locus Ceruleus Neurons in Behaving Monkeys. In *Advances in Pharmacology* (Bd. 42, S. 740–744). Elsevier. https://doi.org/10.1016/S1054-3589(08)60854-6

Ramos, B. P., Colgan, L., Nou, E., Ovadia, S., Wilson, S. R., & Arnsten, A. F. T. (2005). The Beta-1 Adrenergic Antagonist, Betaxolol, Improves Working Memory Performance in Rats and Monkeys. *Biological Psychiatry*, *58*(11), 894–900. https://doi.org/10.1016/j.biopsych.2005.05.022

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Hrsg.), *Classical conditioning II: current research and theory.* (S. 64–99). Appleton Century Crofts.

Reynolds, J. N. J., & Wickens, J. R. (2002). Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks*, *15*(4–6), 507–521. https://doi.org/10.1016/S0893-6080(02)00045-X

Riggall, A. C., & Postle, B. R. (2012). The Relationship between Working Memory Storage and Elevated Activity as Measured with Functional Magnetic Resonance Imaging. *The Journal of Neuroscience*, *32*(38), 12990–12998. https://doi.org/10.1523/JNEUROSCI.1892-12.2012

Roesch, M. R., Calu, D. J., & Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nature Neuroscience*, *10*(12), 1615–1624. https://doi.org/10.1038/nn2013

Roozendaal, B., Okuda, S., De Quervain, D. J.-F., & McGaugh, J. L. (2006). Glucocorticoids interact with emotion-induced noradrenergic activation in influencing different memory functions. *Neuroscience*, *138*(3), 901–910. https://doi.org/10.1016/j.neuroscience.2005.07.049

Rudebeck, P. H., Behrens, T. E., Kennerley, S. W., Baxter, M. G., Buckley, M. J., Walton, M. E., & Rushworth, M. F. S. (2008). Frontal Cortex Subregions Play Distinct Roles in Choices between Actions and Stimuli. *Journal of Neuroscience*, *28*(51), 13775–13785. https://doi.org/10.1523/JNEUROSCI.3541-08.2008

Rushworth, M. F., Kolling, N., Sallet, J., & Mars, R. B. (2012). Valuation and decision-making in frontal cortex: One or many serial or parallel systems? *Current Opinion in Neurobiology*, *22*(6), 946–955. https://doi.org/10.1016/j.conb.2012.04.011

Scatton, B., Claustre, Y., Cudennec, A., Oblin, A., Perrault, G., Sanger, D., & Schoemaker, H. (1997). Amisulpride: From animal pharmacology to therapeutic action. *International Clinical Psychopharmacology*, *12*, 29–36.

Schönberg, T., Daw, N. D., Joel, D., & O'Doherty, J. P. (2007). Reinforcement Learning Signals in the Human Striatum Distinguish Learners from Nonlearners during Reward-Based

Decision Making. *The Journal of Neuroscience*, *27*(47), 12860–12867. https://doi.org/10.1523/JNEUROSCI.2496-07.2007

Schuck, N. W., & Niv, Y. (2019). Sequential replay of nonspatial task states in the human hippocampus. *Science*, *364*(6447), eaaw5181. https://doi.org/10.1126/science.aaw5181

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, *275*, 1593–1599.

Schulz, E., & Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. *Current Opinion in Neurobiology*, *55*, 7–14. https://doi.org/10.1016/j.conb.2018.11.003

Schwabe, L., Hoffken, O., Tegenthoff, M., & Wolf, O. T. (2011). Preventing the Stress-Induced Shift from Goal-Directed to Habit Action with a -Adrenergic Antagonist. *Journal of Neuroscience*, *31*(47), 17317–17325. https://doi.org/10.1523/JNEUROSCI.3304-11.2011

Schwabe, L., Joëls, M., Roozendaal, B., Wolf, O. T., & Oitzl, M. S. (2012). Stress effects on memory: An update and integration. *Neuroscience & Biobehavioral Reviews*, *36*(7), 1740–1749. https://doi.org/10.1016/j.neubiorev.2011.07.002

Schwabe, L., & Schächinger, H. (2018). Ten years of research with the Socially Evaluated Cold Pressor Test: Data from the past and guidelines for the future. *Psychoneuroendocrinology*, *92*, 155–161. https://doi.org/10.1016/j.psyneuen.2018.03.010

Schwabe, L., & Wolf, O. T. (2009). Stress Prompts Habit Behavior in Humans. *Journal of Neuroscience*, *29*(22), 7191–7198. https://doi.org/10.1523/JNEUROSCI.0979-09.2009

Seamans, J. K., Gorelova, N., Durstewitz, D., & Yang, C. R. (2001). Bidirectional Dopamine Modulation of GABAergic Inhibition in Prefrontal Cortical Pyramidal Neurons. *The Journal of Neuroscience*, *21*(10), 3628–3638. https://doi.org/10.1523/JNEUROSCI.21-10-03628.2001

Seamans, J. K., & Yang, C. R. (2004). The principal features and mechanisms of dopamine modulation in the prefrontal cortex. *Progress in Neurobiology*, *74*(1), 1–58. https://doi.org/10.1016/j.pneurobio.2004.05.006

Shafiei, N., Gray, M., Viau, V., & Floresco, S. B. (2012). Acute Stress Induces Selective Alterations in Cost/Benefit Decision-Making. *Neuropsychopharmacology*, *37*(10), 2194–2209. https://doi.org/10.1038/npp.2012.69

Sharp, M. E., Foerde, K., Daw, N. D., & Shohamy, D. (2016). Dopamine selectively remediates 'model-based' reward learning: A computational approach. *Brain*, *139*(2), 355–364. https://doi.org/10.1093/brain/awv347

Shen, W., Flajolet, M., Greengard, P., & Surmeier, D. J. (2008). Dichotomous Dopaminergic Control of Striatal Synaptic Plasticity. *Science*, *321*(5890), 848–851. https://doi.org/10.1126/science.1160575

Sloman, S. A. (1996). The Empirical Case for Two Systems of Reasoning. *Psychological Bulletin*, *119*(1).

Smittenaar, P., FitzGerald, T. H. B., Romei, V., Wright, N. D., & Dolan, R. J. (2013). Disruption of Dorsolateral Prefrontal Cortex Decreases Model-Based in Favor of Model-free Control in Humans. *Neuron*, *80*(4), 914–919. https://doi.org/10.1016/j.neuron.2013.08.009

Sreenivasan, K. K., Curtis, C. E., & D'Esposito, M. (2014). Revisiting the role of persistent neural activity during working memory. *Trends in Cognitive Sciences*, *18*(2), 82–89. https://doi.org/10.1016/j.tics.2013.12.001

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2018). The hippocampus as a predictive map. *Nature Neuroscience*, *20*(11), 23.

Stalnaker, T. A., Cooch, N. K., McDannald, M. A., Liu, T.-L., Wied, H., & Schoenbaum, G. (2014). Orbitofrontal neurons infer the value and identity of predicted outcomes. *Nature Communications*, *5*(1). https://doi.org/10.1038/ncomms4926

Stanovich, K. E., & West, R. F. (2000). Advancing the rationality debate. *Behavioral and Brain Sciences*, *23*(5), 701–717. https://doi.org/10.1017/S0140525X00623439

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2004). Matching Behavior and the Representation of Value in the Parietal Cortex. *Science*, *304*(5678), 1782–1787. https://doi.org/10.1126/science.1094765

Summerfield, C., & Koechlin, E. (2008). A Neural Representation of Prior Information during Perceptual Inference. *Neuron*, *59*(2), 336–347. https://doi.org/10.1016/j.neuron.2008.05.021

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

Tervo, D. G. R., Proskurin, M., Manakov, M., Kabra, M., Vollmer, A., Branson, K., & Karpova, A. Y. (2014). Behavioral Variability through Stochastic Choice and Its Gating by Anterior Cingulate Cortex. *Cell*, *159*(1), 21–32. https://doi.org/10.1016/j.cell.2014.08.037

Thorndike, E. (1927). The Law of Effect. *The American Journal of Psychology*, *39*(1/4), 212–222.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, *185*(4157), 1124–1131.

Ulrich-Lai, Y. M., & Herman, J. P. (2009). Neural regulation of endocrine and autonomic stress responses. *Nature Reviews Neuroscience*, *10*(6), 397–409. https://doi.org/10.1038/nrn2647

Vaessen, T., Hernaus, D., Myin-Germeys, I., & Van Amelsvoort, T. (2015). The dopaminergic response to acute stress in health and psychopathology: A systematic review. *Neuroscience & Biobehavioral Reviews*, *56*, 241–251. https://doi.org/10.1016/j.neubiorev.2015.07.008

Viviani, R., Graf, H., Wiegers, M., & Abler, B. (2013). Effects of amisulpride on human resting cerebral perfusion. *Psychopharmacology*, *229*(1), 95–103. https://doi.org/10.1007/s00213-013-3091-z

Vogel, S., Fernández, G., Joëls, M., & Schwabe, L. (2016). Cognitive Adaptation under Stress: A Case for the Mineralocorticoid Receptor. *Trends in Cognitive Sciences*, *20*(3), 192–203. https://doi.org/10.1016/j.tics.2015.12.003

Warren, C. M., Wilson, R. C., van der Wee, N. J., Giltay, E. J., van Noorden, M. S., Cohen, J. D., & Nieuwenhuis, S. (2017). The effect of atomoxetine on random and directed exploration in humans. *PLOS ONE*, *12*(4), e0176034. https://doi.org/10.1371/journal.pone.0176034

Waterhouse, B. D., Moises, H. C., Yeh, H. H., & Woodward, D. J. (1982). Norepinephrine enhancement of inhibitory synaptic mechanisms in cerebellum and cerebral cortex: Mediation by beta adrenergic receptors. *Journal of Pharmacology and Experimental Therapeutics*, *221*, 495–506.

Wickens, J. R., Horvitz, J. C., Costa, R. M., & Killcross, S. (2007). Dopaminergic Mechanisms in Actions and Habits. *Journal of Neuroscience*, *27*(31), 8181–8183. https://doi.org/10.1523/JNEUROSCI.1671-07.2007

Wiehler, A., Chakroun, K., & Peters, J. (2021). Attenuated Directed Exploration during Reinforcement Learning in Gambling Disorder. *The Journal of Neuroscience*, *41*(11), 2512–2522. https://doi.org/10.1523/JNEUROSCI.1607-20.2021

Wikenheiser, A. M., & Schoenbaum, G. (2016). Over the river, through the woods: Cognitive maps in the hippocampus and orbitofrontal cortex. *Nature Reviews Neuroscience*, *17*(8), 513–523. https://doi.org/10.1038/nrn.2016.56

Wilson, R. C., Bonawitz, E., Costa, V. D., & Ebitz, B. (2021). Balancing exploration and exploitation with information and randomization. *Current Opinion in Behavioral Sciences*, *38*, 49–56.

Wittmann, B. C., Daw, N. D., Seymour, B., & Dolan, R. J. (2008). Striatal Activity Underlies Novelty-Based Choice in Humans. *Neuron*, *58*(6), 967–973. https://doi.org/10.1016/j.neuron.2008.04.027

Woodcock, E. A., Greenwald, M. K., Khatib, D., Diwadkar, V. A., & Stanley, J. A. (2019). Pharmacological stress impairs working memory performance and attenuates dorsolateral prefrontal cortex glutamate modulation. *NeuroImage*, *186*, 437–445. https://doi.org/10.1016/j.neuroimage.2018.11.017

Wunderlich, K., Smittenaar, P., & Dolan, R. J. (2012). Dopamine Enhances Model-Based over Model-Free Choice Behavior. *Neuron*, *75*(3), 418–424. https://doi.org/10.1016/j.neuron.2012.03.042

Yoo, A. H., & Collins, A. G. E. (2022). How Working Memory and Reinforcement Learning Are Intertwined: A Cognitive, Neural, and Computational Perspective. *Journal of Cognitive Neuroscience*, *34*(4), 551–568. https://doi.org/10.1162/jocn_a_01808

Zajkowski, W. K., Kossut, M., & Wilson, R. C. (2017). A causal role for right frontopolar cortex in directed, but not random, exploration. *eLife*, *6*. https://doi.org/10.7554/eLife.27430

Appendix A: Study 1

Cremer, A., Kalbe, F., Gläscher, J., & Schwabe, L. (2021). Stress reduces both model-based and model-free neural computations during flexible learning. *NeuroImage*, *229*, 117747. https://doi.org/10.1016/j.neuroimage.2021.117747

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Stress reduces both model-based and model-free neural computations during flexible learning

Anna Cremer[a], Felix Kalbe[a], Jan Gläscher[b,#], Lars Schwabe[a,#,*]

[a] Department of Cognitive Psychology, Institute of Psychology, Universität Hamburg, Hamburg 20146, Germany
[b] Institute for Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Hamburg 20246, Germany

ARTICLE INFO

Keywords:
Stress
Cognitive flexibility
Decision-making
Reinforcement learning

ABSTRACT

Stressful events are thought to impair the flexible adaptation to changing environments, yet the underlying mechanisms are largely unknown. Here, we combined computational modeling and functional magnetic resonance imaging (fMRI) to elucidate the neurocomputational mechanisms underlying stress-induced deficits in flexible learning. Healthy participants underwent a stress or control manipulation before they completed, in the MRI scanner, a Markov decision task, frequently used to dissociate model-based and model-free contributions to choice, with repeated reversals of reward contingencies. Our results showed that stress attenuated the behavioral sensitivity to reversals in reward contingencies. Computational modeling further indicated that stress specifically affected the use of value computations for subsequent action selection. This reduced application of learned information on subsequent behavior was paralleled by a stress-induced reduction in inferolateral prefrontal cortex activity during model-free computations. For model-based learning, stress decreased specifically posterior, but not anterior, hippocampal activity, pointing to a functional segregation of model-based processing and its modulation by stress along the hippocampal longitudinal axis. Our findings shed light on the mechanisms underlying deficits in flexible learning under stress and indicate that, in highly dynamic environments, stress may hamper both model-based and model-free contributions to adaptive behavior.

Stressful events are a powerful modulator of learning and memory (Diamond et al., 2007; Luksys and Sandi, 2011; Lupien et al., 2009; Roozendaal et al., 2009; Schwabe et al., 2012). In particular, stress is thought to render learning and memory rather rigid, thus impairing the flexible adaptation to changing environments (Raio et al., 2017; Schwabe and Wolf, 2013; Wirz et al., 2018; Schwabe et al., 2013). Although such deficits in flexible learning under stress have far-reaching implications, not only for educational and clinical contexts (de Quervain et al., 2017; Goldfarb and Sinha, 2018; Goodman et al., 2012; Vogel and Schwabe, 2016), the exact mechanisms underlying the stress-induced impairments in flexible learning are still largely unclear.

Successful adaptation to dynamic environments depends on the complex interplay of at least two systems: (i) a reflective or goal-directed system that involves the consideration of prospective future courses of action and their consequences and (ii) a reflexive or habitual system that is guided by the retrospective experience of good and bad outcomes (Balleine and O'doherty, 2010; Sloman, 1996). Accumulating evidence from human and rodent studies shows that stress and stress hormones may bias the balance of these systems and favor habitual over goal-directed behavior (Braun and Hauber, 2013; Gourley et al., 2012; Schwabe et al., 2012; Schwabe and Wolf, 2009, 2011; Schwabe et al., 2012b). Computationally, goal-directed and habitual forms of behavioral control are assumed to overlap to some degree with model-based and model-free reinforcement learning systems (Dolan and Dayan, 2013). Within this framework, learning can be defined as the identification of a value function that selects the most rewarding options in the current environment. Therefore, the value function links the previous value of the options available with rewards that can be expected in the future. This results in a policy that maps different environments to action probabilities and therefore determines which actions are selected in each state (Gershman and Uchida, 2019). Specifically, a central aspect in both model-based and model-free learning is the computation of prediction error signals to update the value function. Therefore, previous experiences are used to form predictions, which are then updated by comparing the predicted outcome of an option to the actual outcome.

While a model-based policy acquires a cognitive map of the task structure (i.e., how different environments are linked to each other) and uses this to predict the most advantageous course of action, the model-free system encodes values by trial and error and uses the reward history to guide behavior (Daw et al., 2005, 2011; Gläscher et al., 2010).

[*] Corresponding author.
E-mail address: Lars.Schwabe@uni-hamburg.de (L. Schwabe).
[#] these authors contributed equally

https://doi.org/10.1016/j.neuroimage.2021.117747
Received 27 August 2020; Received in revised form 4 January 2021; Accepted 9 January 2021
Available online 14 January 2021

On the neural level, model-based processing is thought to rely on posterior inferior parietal as well as lateral prefrontal regions (Gläscher et al., 2010) and, as shown more recently, on hippocampal areas (Pfeiffer and Foster 2013; Garvert et al., 2017; Miller et al., 2017; Stachenfeld et al., 2017). Model-free learning, in turn, is assumed to be driven by prediction error signals of midbrain dopamine neurons mapping the difference between the actual and expected reward at a particular state and depends mainly on the ventral striatum (Bayer and Glimcher, 2005; Haruno and Kawato, 2006; McClure et al., 2003; O'Doherty et al., 2003). In terms of the flexible adaption to changes in the environment, we identified the medial prefrontal cortex (mPFC) as a potential key player, since it is linked to essential features of flexible learning (Nee et al., 2011). In particular, the mPFC is thought to be implicated in the anticipation of values of currently available actions (Aarts et al., 2008), the representation of possible outcomes (Brown, 2009), the association between actions and outcomes (Oliveira et al., 2007), error detection processes during contingency changes (Zarr and Brown, 2016), and the computation of likely action outcomes (Alexander and Brown, 2011; Croxson et al., 2009).

The computational conceptualization of reflexive and reflective systems of behavioral control in terms of model-based and model-free processing provided valuable insights into the mechanisms underlying each of these systems as well as their interplay. First behavioral studies suggested that acute stress may affect behavioral flexibility in general (Plessow et al., 2011; Schwabe and Wolf, 2011) and the contributions of model-based and model-free processes to aversive learning or learning from negative outcomes in particular (Park et al., 2017; Raio et al., 2017). However, how stress changes the contributions of model-based and model-free systems to flexible learning in a highly volatile environment and, in particular, the neural mechanisms underlying stress-induced alterations in model-based and model-free processing are largely unknown.

In the present experiment, we combined computational modeling and functional magnetic resonance imaging (fMRI) to elucidate the neurocomputational mechanisms underlying stress-induced deficits in flexible learning. Therefore, healthy participants first underwent a standardized stress or control procedure before they completed a two-step Markov decision task in the MRI scanner. This task allows a dissociation of model-based and model-free contributions to behavior (Daw et al., 2011) and requires two subsequent decisions which can ultimately lead to a reward. To explicitly probe the flexibility of learning, we used a modified version of this task that included repeated reversals of reward contingencies. Here, flexible learning was expressed as the ability to detect a reversal and adapt the choice behavior accordingly. We assumed that task performance would rely on both model-based and model-free computations and that stress would reduce their recruitment during learning. Because previous findings suggested that individuals with low working memory capacity were more susceptible to detrimental stress effects on model-based learning strategies than participants with high working memory capacity (Otto et al., 2013), we further included an n-back test to probe participants' baseline working memory performance.

1. Materials and methods

1.1. Participants and experimental design

Sixty-eight healthy volunteers participated in this experiment. Based on previous studies from our lab that reported effect sizes of Cohen's d from 0.66 to 0.98 for similar research questions (Schwabe and Wolf, 2009, 2012), we expected a medium to large effect of stress on flexible learning of Cohen's $d = 0.7$. A power analysis using G*power (Faul et al., 2007) indicated that using a two-tailed independent t-test with alpha = 0.05, a sample of 68 participants is required to detect such a medium-sized effect with a power of 0.80. All participants were right-handed, had normal or corrected-to-normal vision and were screened for possible MRI contraindications. Individuals with a current medical condition, current medication intake or lifetime history of any neurological or psychiatric disorders were excluded from participation. Moreover, we excluded smokers and women taking hormonal contraceptives as both can affect the stress response (Kirschbaum et al., 1999; Rohleder and Kirschbaum, 2006). Participants were asked not to drink coffee or other caffeinated beverages and not to do any exercise on the day of the experiment. In addition, they should not eat or drink anything except water 2 h before the appointment. All participants provided written informed consent before the beginning of testing and received a moderate monetary compensation. The study protocol was approved by the local ethics committee. Ten participants had to be excluded from the analysis because of excessive head movement (mean displacement > 5 mm) in the MRI ($n = 4$), because they missed more than 30% of the trials ($n = 3$) or because they chose the same action in more than 95% of the trials ($n = 3$), thus leaving a final sample of 58 participants (17 men and 12 women in each of the two groups, age 18–34, mean = 24.6, SD = 3.5, no age difference between groups, t(57) = 0.73, $p = 0.47$). Participants were pseudorandomly assigned to the stress and control groups, in order to achieve an identical number of men and women per group.

1.2. Stress induction

In order to control for the diurnal rhythm of the stress hormone cortisol, all testing took place in the afternoon and early evening, with the time of testing being counterbalanced across groups. Participants of the stress group underwent the Trier Social Stress Test (TSST; Kirschbaum et al., 1993), a standardized paradigm in experimental stress research that is known to activate both the autonomic nervous system and the hypothalamus-pituitary-adrenal axis. In brief, the TSST simulates a 15-min job interview, including a public speech about the participant's eligibility for a job tailored to his/her interests and a mental arithmetic task. During both tasks, participants were videotaped and evaluated by two rather cold, non-reinforcing committee members (1 man, 1 woman), dressed in white lab coats. In the control condition, participants spoke about a topic of their choice followed by a simple arithmetic task (counting forwards in steps of 15), without committee or video recordings.

To evaluate the successful stress induction through the TSST, subjective and physiological measurements were taken at several time points across the experiment (see Fig. 1). Baseline was assessed 10 min after the start of the appointment, so that the subjects were able to acclimatize to the situation. Directly after the TSST/control manipulation, participants rated the difficulty, stressfulness, and unpleasantness of the experimental treatment on a scale from 0 ("not at all difficult/stressful/unpleasant") to 100 ("very difficult/stressful/unpleasant"). Blood pressure and pulse were measured at baseline, during the TSST, directly after the TSST and after the fMRI scanning session using a digital blood pressure device (OMRON model M500 (HEM-7321-D); Healthcare Europe BV, Hoofddorp, The Netherlands) with a cuff applied around the right upper arm, when subjects were standing. Finally, we collected saliva samples using Salivette® collection devices (Sarstedt) at baseline, 18 min after stressor onset (shortly before the learning task started), and after each block of the Markov decision task (i.e., 40, 60 and 90 min after the treatment, Fig. 1). Saliva samples were stored at −18 °C until the end of data collection, when we analyzed saliva cortisol concentrations using a luminescence assay (IBL, Germany).

1.3. Markov decision task

Twenty minutes after the beginning of the stress/control manipulation, when stress-induced cortisol concentrations were expected to peak, participants performed a modified version of a two-step Markov decision task in the MRI scanner. This task was designed to dissociate between model-based and model-free learning mechanisms (Daw et al., 2011). Each trial consisted of two successive stages, in each of which

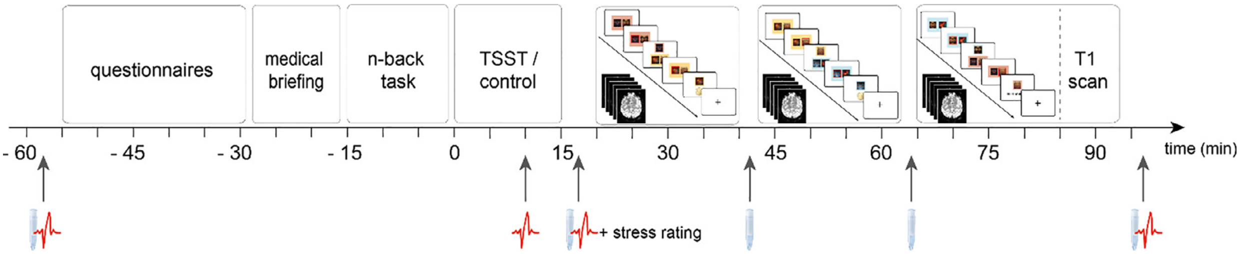


Fig. 1. Experimental procedure. Stress was induced by the Trier Social Stress Test (TSST). Before the stress/control treatment, participants completed several questionnaires and performed an n-back task. After the stress/control procedure, participants completed three blocks of a modified Markov decision task (MDT) in the MRI scanner. Stress reactivity was assessed by subjective and physiological measures (salivary cortisol, blood pressure, pulse), which were taken at several time points across the experiment.

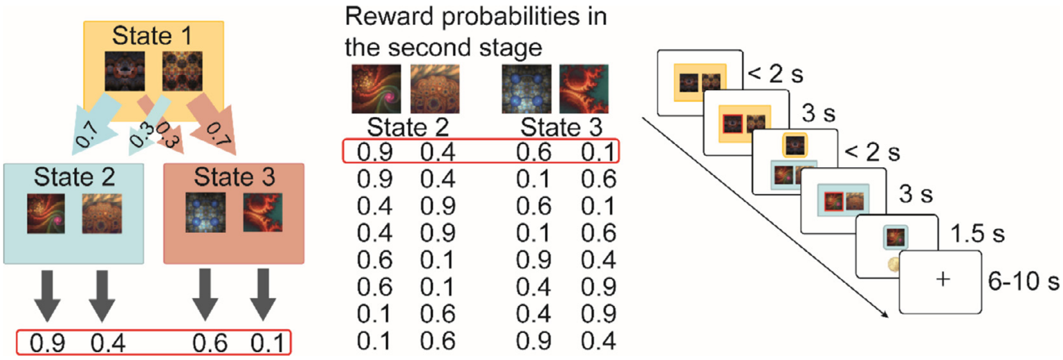


Fig. 2. Experimental task. Left: State transition structure. Each first stage (state 1) action is predominantly associated with one or the other second stage states (state 2 and state 3), and leads there in 70% of the time. The different states were marked by differently colored boxes. Reward probabilities in the second stage undergo frequent reversals. Middle: Reversal patterns. Reward probabilities of 0.4 and 0.9, and 0.1 and 0.6 stay together in one state. Out of these possible combinations, six patterns (two per block) occurred over the course of the experiment per participant. Right: Timeline of events per trial. A first stage choice between two options leads to a second stage choice which is probabilistically reinforced with monetary rewards.

the participant had to choose between two options ("left" or right"), represented by fractal images (Fig. 2). The first stage decision (state 1) led to one of two possible states in the second stage (state 2 and state 3), requiring another choice between two fractals, which were associated with different probabilities of receiving a monetary reward. Each first stage option was predominantly (70%) associated with one or the other second stage state. The choice options in the second stage led to a monetary reward with a probability of 0.9, 0.6, 0.4 and 0.1, with the probabilities of 0.9 and 0.4 being paired in one state ("good state"), and 0.6 and 0.1 in the other ("bad state"). A win was depicted by a 10 cents coin, otherwise "no reward" appeared on the screen. Upon each state, participants had 2 s to submit their choice on an MRI compatible button box. If they failed to enter a choice within this time window, the trial was aborted and the next trial started. Trials were separated by an inter trial interval of randomized length, between 6 and 10 s (Fig. 2, right).

Whether the transition structure is included in the decision or not provides insights into the engaged learning strategy. While the model-free learner evaluates actions retrospectively by repeating previously rewarded choices, the model-based learner also takes the task structure into account. Consider a first stage choice that led to a second state via a rare transition, followed by a second state choice that led to a reward. A purely model-free agent would repeat the action because it was rewarded. A purely model-based learner, however, would switch to the other first stage option because it takes into account that the previous first stage action only leads to the rewarded second stage state via a rare transition. Thus, first stage decisions provide the opportunity to determine the extent to which model-based vs. model-free computations contribute to decisions. In order to explicitly test the flexibility of learning, we modified the original task by introducing repeated reversals of reward contingencies (Fig. 2), requiring the flexible adaptation of behavior.

Participants performed 202 trials, distributed over three blocks (70 / 66 / 66 trials) and separated by breaks. In order to explicitly test the flexibility of learning, defined as adaptation to a changing environment, we modified the original task by introducing repeated reversals of reward contingencies. Specifically, the reward probabilities associated with second stage actions were reversed twice per block, fixed at trial numbers 27, 49, 93, 115, 159 and 181. To ensure that reversals are detectable despite the probabilistic reward structure, the reversals only take place within one of the two second stage states. Fig. 2 (middle) shows all possible reversal combinations. Note that the three blocks were separated by a short break in which the subjects were briefly moved out of the scanner to collect saliva samples. The experimenter then placed the Salivette in the participant's mouth using sterile plastic tweezers, paralleled by the instruction to move as little as possible. We applied the same criteria for the movement parameters between the blocks as during the task, i.e. excluding participants with a mean displacement > 5 mm. To make sure that participants did not continue to apply their previously learned contingencies, each block began with a new stimulus allocation. That is, the same six fractals were randomly assigned to the three states. Likewise, the background colors of the states were reassigned. The second stage reward probabilities were randomly attached to the new second stage stimuli. The assignment of colors and stimuli to the states was counterbalanced across participants. The stimulus pairs within the states stayed the same within one block, and so did the background colors of the states and the transitions between first and second stage. The location of the two options in one state was randomized from trial to trial to ensure that the participants learned stimulus – state and stimulus – reward contingencies rather than the stimulus position.

Participants were instructed that they had to make two decisions in a row in each trial, with the second decision possibly leading to a reward and that the aim of the task was to gain as many rewards as possible. They were told that the first decision was not directly associated with the reward, but that it leads to one of two possible states, in which they again have the choice between two pictures. They were also instructed that each first stage option was primarily associated with one or the other state in the second stage, but not with which one. Further, the instructions stated that within the two second stage states, each picture leads with a certain probability to a reward and in both states there is a slightly better and a slightly less favorable option and that they should find the best option. Most importantly, participants were informed that this option would change several times throughout the experiment and that they should detect the changes and adapt their behavior accordingly. Lastly, they learned that none of the images would lead to a reward in all trials, so that it is possible that they do not get a reward for an answer that has been correct many times before - but that this does not necessarily mean that a reversal took place. The exact task instructions that participants received are provided in the supplemental material.

Before the TSST or control manipulation, participants performed a brief training session for the learning task (out of the MRI scanner). The training consisted of three parts with 10 trials each, introducing the task iteratively. In each phase, the trial structure was the same as in the experimental task. In the first part, the participants should find out which image would lead to a reward with one of the four images being rewarded while the other three were not. The second part was identical, except that the rewarded picture changed at some point. The participants were informed about the reversal and were instructed to adapt their behavior. The third phase was identical to the experimental task. We used the same stimuli and transition structure as in the experiment, the image position and colors of the states were randomized. All three phases had a fixed number of trials, no learning criterion was applied.

1.4. Working memory assessment

Because previous research suggested that the influence of stress on the control of learning may be moderated by the individual working memory capacity (Otto et al., 2013), we measured working memory using an n-back task (Kirchner, 1958) before participants underwent the stress or control manipulation. Participants were presented a random sequence of one-digit numbers from “0” to “9” and asked to indicate via button press (“yes” or “no”) whether the currently presented number was the same as the one presented n-trials before. Participants received 10 stimulus blocks in total (2 practice blocks with feedback and 8 experimental blocks without feedback), in which working memory load varied by alternately using a 2-back and a 3-back condition. Each block consisted of 24 stimulus trials. Stimuli were displayed for 500 ms and responses were recorded within 1500 ms, followed by 2000 ms fixation cross.

1.5. Behavioral data analyses

To test whether the TSST successfully induced stress, data on subjective ratings, vital signs, and salivary cortisol were analyzed using mixed-design ANOVAs with the between-subjects factor treatment and the within-subjects factor time after stress/control manipulation onset. T-tests were used to investigate post-hoc group differences in these measures. Learning performance was quantified by the proportion of first stage choices for the option that led predominantly, with a probability of 0.7, to the second stage state with the overall higher probability to obtain a reward. Likewise, the proportion of choices for the option with the higher reward probability (either 0.9 or 0.6) in the second stage reflected successful learning. We further computed the sensitivity to detect changing contingencies as a difference index between the mean number of advantageous choices in the four trials before a reversal relative to the four trials after a reversal. We chose this number of trial before and after a reversal to ensure that the participants had enough trials to learn the contingencies and to specifically capture the reversal related behavior. The results remained the same if we used, for instance, 5 trials before/after a reversal instead. In order to identify the model-based and model-free contributions to behavior and whether these contributions differed between the stress and control groups, we used a mixed design ANOVA with the between-subjects factor treatment (stress vs. control manipulation) and the within-subject factors reward (rewarded vs. not rewarded) and transition (common vs. rare). Further, we performed a mixed-effects logistic regression to explain the first stage choice on each trial. First stage choice was coded as stay vs. switch and was explained as a function of previous trial’s outcome (rewarded or not rewarded) and previous trial’s transition type (common or rare). Within-subject factors (the intercept, main effects of reward and transition, and their interaction) were taken as random effects across subjects, and estimates and statistics reported at the population level. The experimental treatment (stress vs. control) was taken as a fixed effect.

We also performed exploratory analyses to test whether anxiety, depression, chronic stress or working memory capacity influence the susceptibility to stress effects on flexible learning. We tested whether these measures correlated with the sensitivity index or the model parameters. Additionally, we subdivided the stress group and the control group based on a median split on these measures, and analyzed whether individuals with particular high or low scores differed in their behavior around the reversals by using a mixed design ANOVA with the between-subjects factors treatment (stress vs. control manipulation) and level (high vs. low) and the within-subject factor time (pre reversal vs. post reversal). All analyses were performed in R (R Core Team, 2019). Greenhouse-Geisser correction was applied when sphericity was violated. Logistic regressions were conducted as mixed-effects models and were performed using the lme4 package (Pinheiro and Bates, 2000).

1.6. Computational modeling

We used reinforcement learning models to dissociate model-free and model-based contributions to subject’s trial by trial choices. We fit choice behavior to a dual-system reinforcement learning model which includes both model-free and model based learning strategies, assuming that choices derive from a weighted combination of both model-free and model-based value computations (Daw et al., 2011; Gläscher et al., 2010). Therefore, the algorithms learn a value function $Q(s,a)$ for each of the stimulus-action pairs in the two stages (three states, first stage: s_A, second stage: s_B and s_C; each with two actions). On trial t, the first stage state (always s_A) is followed by the first stage action which leads to the second stage state (s_B or s_C). The second stage action a_2 is probabilistically connected to a reward $r_{2,t}$. At each stage i of each trial t, the value for the visited state-action pair $Q(s,a)$ was updated according to both a model-free and a model-based algorithm.

Model-free values were computed with a SARSA (λ) temporal difference algorithm. As stated before, model-free choices derive from repeating previously rewarded actions. In the first trial, each state-action pair (s, a) at stage i and trial t has a Q-value of zero. In each following trial $t+1$ the value for the visited state-action pair $Q_{MF}(s_{i,\,t+1},\ a_{i,\,t+1})$ is updated based on whether the particular pairing was rewarded in the previous trial t. Therefore, the general form of the model-free value update for chosen stimulus-action pair is:

$$Q_{MF}\left(s_{i,\,t+1},\ a_{i,\,t+1}\right) = Q_{MF}\left(s_{i,\,t},\ a_{i,\,t}\right) + \alpha_i \delta_{i,\,t} \tag{1}$$

where

$$\delta_{i,\,t} = r_{i,\,t} + Q_{MF}\left(s_{i+1,\,t},\ a_{i+1,\,t}\right) - Q_{MF}\left(s_{i,\,t},\ a_{i,\,t}\right) \tag{2}$$

δ refers to the reward prediction error and α indicates the learning rates. However, this general form of value update and prediction error is narrowed down in the different stages of the task, which is explained next.

The prediction error is different for the two stages of the task. Since $r_{1,t}$ is always zero, the prediction error at the first stage is driven by the value of the selected second stage action:

$$\delta_{1,\ t} = Q_{MF}\left(s_{2,\ t},\ a_{2,\ t}\right) - \ Q_{MF}\left(s_{1,\ t},\ a_{1,\ t}\right) \tag{3}$$

This prediction error $\delta_{1,\ t}$ is used to update $Q_{MF}(s_{1,\ t},\ a_{1,\ t})$ immediately after the first choice has been made:

$$Q_{MF}\left(s_{1,\ t+1},\ a_{1,\ t+1}\right) = Q_{MF}\left(s_{1,\ t},\ a_{1,\ t}\right) + \ \alpha_1\delta_{1,\ t} \tag{4}$$

Since there is no third stage, the second stage prediction error is driven by the reward $r_{2,t}$:

$$\delta_{2,\ t} = r_{2,t} - \ Q_{MF}\left(s_{2,\ t},\ a_{2,\ t}\right) \tag{5}$$

This prediction error $\delta_{2,\ t}$ at the second stage is used to update the first and second stage model-free values, once the reward information of the outcome has become available.

$$Q_{MF}\left(s_{2,\ t+1},\ a_{2,\ t+1}\right) = Q_{MF}\left(s_{2,\ t},\ a_{2,\ t}\right) + \ \alpha_2\delta_{2,\ t} \tag{6}$$

$$Q_{MF}\left(s_{1,\ t+1},\ a_{1,\ t+1}\right) = Q_{MF}\left(s_{1,\ t},\ a_{1,\ t}\right) + \ \alpha_1\lambda\delta_{2,\ t} \tag{7}$$

Note that this update uses the already updated $Q_{MF}(s_{1,\ t},\ a_{1,\ t})$ from above, thus constituting a second update of first stage values.

The learning rates a_1 and a_2, estimated for both stages, control how much the Q-value is updated by the prediction error and therefore indicate to what extent newly acquired information overwrites old information. The learning rates are constrained between 0 and 1 with an α parameter = 0 indicating no learning and $\alpha = 1$ indicating the agent considers only the most recent information. Further, at the end of each trial the eligibility parameter λ (range 0 to 1) modulates a_1 in the second update in light of the reward information that has become available at the end of the trial . Higher values of lambda indicate more reliance to further back states and actions. In other words, λ performs a down-weighting of the first stage action based on the temporal distance from the current trial. Both the first- and second stage Q_{MF} values are updated at the second stage, with the first stage values receiving the prediction error values that were decayed by λ (see supplement in Daw et al. (2011) for details).

The model-based agent learns cumulative state-action values with a FORWARD algorithm. As described before, the model-based learner's decisions are not only determined by the reward, but also include the path that lead to the second stage's state, i.e. whether the transition was common or rare. Specifically, the algorithm computes a transition function for the first stage state-action pairs and then combines it with the second stage's reward predictions. Referring to our experimental task, this means that a model-based learner would first consider which first stage action leads to which second stage state, and then learn the reward values for the second stage actions. At the first stage, the transition function T contains the information of which first stage action maps to which second stage state. Note that in our model, the transition structure with common and rare transitions leading to 70 and 30 percent in one of the two states in the second stage was predetermined and not learned by the model (see below for a test of this supposition). At the second stage, Q_{MB} values are calculated similar to the Q_{MF} values: comparing the actual outcome of the visited state with the predicted outcome, weighted by the learning rate α (to which extent will the old information be overwritten by the new information) and the eligibility parameter λ (how far is the distance from the current trial). Model-based value expectation depends on the specification of first stage Q-values in terms of Bellman's equation (Sutton and Barto, 1998) using the transition structure P:

$$\begin{aligned}Q_{MB}\left(\mathrm{s}_{A,t+1}, a_{1,t+1}\right) = {} & \mathrm{P}(s_B\ |\mathrm{s}_{A,\ t})\ max\ Q_{MF}\left(s_{B,t},\ a_{2,t}\right) \\ & +P\left(s_C\ \mathrm{s}_A, a_{1,t}\right)\ max\ Q_{MF}\left(s_{C,t},\ a_{2,t}\right)\end{aligned} \tag{8}$$

and is recomputed at each trial, based on the current estimates of the transition probabilities and second stage reward values. Because model-based and model-free algorithms coincide at the second stage, we set $Q_{MB} = \ Q_{MF}$ at this level.

Finally, we assume that behavior derives from a weighted combination of both model-based and model-free value computations. Therefore, we define net action values at the first stage as the weighted sum of model-based and model-free values

$$Q_{net}\left(\mathrm{s}_{A,t+1}, a_{1,t+1}\right) = wQ_{MB}\left(\mathrm{s}_{A,t}, a_{1,t}\right) + (1-w)Q_{MF}\left(\mathrm{s}_{A,t}, a_{1,t}\right), \tag{9}$$

where w is a weighting parameter. This parameter is assumed to be constant across trials, with $w = 0$ reflecting purely model-free value computing and $w = 1$ purely model-based reinforcement learning. The probability of a choice is composed by a softmax for Q_{net} at the first stage:

$$\mathrm{P}(a_{1,\ t} = a\ |\mathrm{s}_{A,\ t}) = \frac{\exp\left(\beta_1\left[Q_{net}\left(s_{1,\ t},\ a\right) + p\ *rep(a)\right]\right)}{\sum_{a'}\exp\left(\beta_1\left[Q_{net}\left(s_{1,\ t},\ a'\right) + p\ *rep(a')\right]\right)} \tag{10}$$

where the inverse temperature parameters β_1 and β_2 indicate the randomness of the choice by specifying the extent to which the values are updated based on the learned information. Temperature parameters are set from 0 to ∞ with lower values indicating more randomness in choice behavior. The stay parameter p, ranging from 0 to 1, captures first-order perseveration in the first stage, together with the indicator function rep that is 1 when the current first stage action is the same as in the previous trial. The stay parameter was omitted at the second stage and hence the softmax is defined as:

$$\mathrm{P}(a_{2,\ t} = a\ |\mathrm{s}_{2,\ t}) = \frac{\exp\left(\beta_2\left[Q_{net}\left(s_{2,\ t},\ a\right)\right]\right)}{\sum_{a'}\exp\left(\beta_2\left[Q_{net}\left(s_{2,\ t},\ a'\right)\right]\right)} \tag{11}$$

In total, the algorithm contains 7 free parameters (a_1, a_2, β_1, β_2, λ, p, w) which were fit separately for each participant using the probabilistic programming language Stan through its MATLAB interface (Carpenter et al., 2017).

With the help of the model-parameters determined for each subject we were able to draw conclusions about the learning strategies used. We conducted group comparisons for each parameter to identify general differences in behavioral tendencies between the stress group and the control group.

To determine different learning strategies in the neuroimaging data we calculated three different prediction errors. Therefore, we extracted each participant's best fitting parameters and reran the task, resulting in model predictions on a trial basis. In addition to the actual prediction with the individual w-parameter, we also created model-based or model-free predictions by again inserting the parameters in the task, but this time not the fitted w-parameter, but with $w = 0$ and $w = 1$, reflecting pure model-free and pure model-based behavior, respectively. Thus, we obtain three predicted datasets for each subject. This allows us to distinguish between model-based and model-free prediction errors for the value update at stage 1.

For Q_{MB}, we set $w = 1$:

$$Q_{MB}\left(\mathrm{s}_{i,t+1}, a_{i,t+1}\right) = 1{*}Q_{MB}\left(\mathrm{s}_{i,t}, a_{i,t}\right) + (1-1)Q_{MF}\left(\mathrm{s}_{i,\ t}, a_{i,t}\right) \tag{12}$$

Likewise, for Q_{MF}, we determine $w = 0$:

$$Q_{MF}\left(\mathrm{s}_{i,t+1}, a_{i,t+1}\right) = 0{*}Q_{MB}\left(\mathrm{s}_{i,t}, a_{i,t}\right) + (1-0)Q_{MF}\left(\mathrm{s}_{i,t}, a_{i,t}\right) \tag{13}$$

These predicted Q-values are used to derive prediction errors (see Eq. (3)):

$$PE_{MB}\left(s_{1,t},\ a_{1,\ t}\right) = Q_{MB}\left(\mathrm{s}_{2,t}, a_{2,t}\right) - Q_{MB}\left(\mathrm{s}_{1,t}, a_{1,t}\right) \tag{14}$$

$$PE_{MF}\left(s_{1,t},\ a_{1,\ t}\right) = Q_{MF}\left(\mathrm{s}_{2,t}, a_{2,t}\right) - Q_{MF}\left(\mathrm{s}_{1,t}, a_{1,t}\right) \tag{15}$$

Finally, we identify the reward prediction error RPE which is calculated when the outcome is presented:

$$RPE\left(s_{2,\ t},\ a_{2,\ t}\right) = r_{2,t} - \ Q_{net}\left(s_{2,\ t},\ a_{2,\ t}\right) \tag{16}$$

Table 1
Model Comparison using WAIC.

Model Name	α	β	p	w	λ	ε	**nParams**	**Control**	**Stress**
Full	**2**	**2**	**1**	**1**	**1**	**0**	**7**	**11,191.15**	**12,056.88**
full + state space	2	2	1	1	1	1	8	11,202.57	12,062.26
no p	2	2	0	1	1	0	6	11,436.80	12,359.40
one α	1	2	1	1	1	0	6	11,214.76	12,072.60
one β	2	1	1	1	1	0	6	11,218.68	12,077.35
no p_one α	1	2	0	1	1	0	5	11,495.35	12,450.99
no p_one β	2	1	0	1	1	0	5	11,542.43	12,433.98
one α_one β	1	1	1	1	1	0	5	11,541.47	12,436.59
no p_one α_one β	1	1	0	1	1	0	4	11,628.99	12,509.57

1.7. Model validation

To validate the model fit, we compared our fully parameterized hybrid model (Daw et al., 2011; Gläscher et al., 2010) to various reduced nested versions. The model should be as complex as necessary to adequately represent behavior, but only as complex as justified by the data. We compared our model to several other models that are simplified by removing different parameters, e.g. a version without the stay bias (p), a version with only one learning rate (α), a version with only one temperature parameter (β), and combinations of these reductions (Table 1). We also explicitly tested whether state space learning played a significant role in task performance because the participants did not know the transition probabilities for common and rare transitions at the beginning of the experiment. We therefore included a mechanism by which a participant can learn the transition probabilities during task execution, which we have used in a prior publication (Gläscher et al., 2010). Transition probabilities are stored in a transition matrix T, which can be learned using a state prediction error (see Gläscher et al., 2010). Specifically, each element of T specifies the probability of the reached second stage state s_2 from the first stage state s_1 via an action a_1 ($T(s_{1,t}, a_{1,t}, s_{2,t})$). In the state space learning model, all transition probabilities are initially set to 0.5 reflecting no prior knowledge about the transitions by the participants. Upon every trial all possible transitions following action $a_{1,t}$ are updated according to the following learning rules:

$$T\left(s_{1,t+1}, a_{1,t+1}, s_{2,t+1}\right) = T\left(s_{1,t}, a_{1,t}, s_{2,t}\right) + \varepsilon(1 - T\left(s_{1,t}, a_{1,t}, s_{2,t}\right) \tag{17}$$

$$T\left(s_{1,t+1}, a_{1,t+1}, \neg s_{2,t+1}\right) = T\left(s_{1,t}, a_{1,t}, \neg s_{2,t}\right) - \varepsilon\ T\left(s_{1,t}, a_{1,t}, \neg s_{2,t}\right) \tag{18}$$

where $T(s_{1,t}, a_1, s_{2,t})$ is the probability of transitions from the first stage state $s_{1,t}$ to the second stage state $s_{2,t}$ using action $a_{1,t}$ on trial t, $T(s_{1,t}, a_1, \neg s_{2,t})$ is the unrealized transition to the other possible second stage state, and ε is the learning rate for state space learning, modeled with an initial uniform Beta(1,1) prior. We think that updating both the realized and the unrealized state transition following the same action $a_{1,t}$ is a reasonable approach given that participants are probably aware (at least in the latter parts of the experiment) that action $a_{1,t}$ could have also resulted in a different transition. All other components of the state space learning model are identical to the full learning model, including the linear weighting of model-free and model-based learning (see equations above). Model comparisons were performed by calculating the widely applicable information criterion (WAIC; Watanabe, 2010) which indicates prediction performance and assesses the quality of a model, relative to the quality of other candidate models by estimating the posterior likelihood, followed by a correction for the effective number of parameters to adjust for overfitting. This approach is often used for comparing models estimated using Markov Chain Monte Carlo sampling as in our case and confirmed that the full model outperformed all competing versions (Table 1).

A fully parameterized hybrid model without a state space learning component fitted subjects' choices best in a model comparison that considers differences in model complexity. Model performance is indicated by the widely applicable information criterion (WAIC), presented separately for the stress group and the control group. Lower values represent a better fit. The full model contains two learning rates (α), two temperature parameters (β), the stay bias (p), the weighting parameter (w) and the eligibility parameter (λ) and was compared to several other models that are simplified by removing different parameters, respectively, or included the state space learning rate ε.

1.8. MRI data acquisition and analysis

Functional imaging was conducted using a 3 T Siemens (Erlangen, Germany) MAGNETOM Prisma scanner, equipped with a 64-channel head coil, to acquire gradient echo T2*-weighted echo-planar-images (EPI) with BOLD contrast. For each of the three functional runs, we collected about 600 vol with the following parameters: 60 slices, slice thickness = 2 mm, flip angle 60%, FOV 224 × 224, repetition time (TR) = 2000 ms, echo time (TE) = 30 ms, voxel size 2.0 mm isotropic. Slice orientation was tilted –30° from the line connecting the anterior and posterior commissure to alleviate signal drop out in the orbitofrontal cortex (Deichmann et al., 2003). We additionally acquired a high-resolution T1-weighted anatomical image (TR = 2.5 s, TE = 2.12 ms, 256 slices, voxel size = 0.8 × 0.8 × 0.9 mm). Preprocessing of functional images was performed with MATLAB and SPM12 (http://www.fil.ion.ucl.ac.uk/spm/). The first five functional images were discarded from the analysis to allow for T1 saturation effects. The remaining functional images were first spatially realigned, then coregistered to the structural image, followed by a normalization to the MNI space. The images were additionally spatially smoothed using a 4 mm full-width half-maximum Gaussian kernel.

Subject-specific design matrices were defined using general linear modeling (GLM) as implemented in SPM12. We entered three regressors coding the average BOLD response at each of the three states (two choice states, one outcome state). The model-derived prediction error signals (model-free prediction error PE_{MF} and model-based prediction error PE_{MB}) were entered as parametric modulators, modeled at the onset of the second stage. We chose this time point because we assume that the relevant learning computations are integrated to a value update when the second decision is required. A parametric regressor coding the received outcomes (reward = +1, no reward = 0) was modeled at the time of the outcome. Model-derived reward prediction errors (RPE) were modeled as parametric modulators on the outcome onsets, because here we are specifically interested in the process of reward-related value updating. Moreover, we subdivided behavioral data into "advantageous-choice-trials" and "disadvantageous-choice-trials", separately for the first and the second stage and entered their onsets into our GLM. For the second-level models, the contrasts of interest "model-based prediction error", "model-free prediction error", "reward prediction error", "reward", "advantageous choice stage 1" and "advantageous choice stage 2" were defined. These difference contrasts were taken to a second-level group two-sample *t*-test, allowing a direct comparison between the stress and control group.

Based on our a-priori hypotheses, analyses were restricted to brain areas that have previously been implicated in model-based and model-free reinforcement learning (Daw et al., 2011; Gläscher et al., 2010; Lee et al., 2014). We used the following anatomical masks from the Harvard-Oxford atlas: putamen, caudate, and hippocampus. In order to test for potential differential involvement of anterior and posterior regions of the hippocampus in model-based and model-free learning under stress, we divided a hippocampal mask along the y-axis into three parts with approximately equal lengths, using the WFU pick-atlas (Lancaster et al., 2000; Maldjian et al., 2003): posterior hippocampus from $Y = -40$ to -30, medial hippocampus from $Y = -29$ to -19, and anterior hippocampus from $Y = -18$ to -4. For a more detailed description see Collin et al. (2015) and Dandolo and Schwabe (2018). Moreover, we used anatomical masks for lateral orbitofrontal cortex from the Montreal atlas and combined the AAL atlas-masks for frontal superior medial, frontal middle and frontal superior to a medial prefrontal cortex mask, as implemented in the WFU PickAtlas Tool (Maldjian et al., 2003). 10 mm spheres centered on the peak voxel of bilateral ventral striatum (left peak: −9 2 8, right peak: 9 5 −8), bilateral insulae (left peak: −30 20 −2, right peak: 33 29 7) and ilPFC (left peak: −54 38 3, right peak: 48 35 −2) were created, because they were previously associated with model-free and model-based learning strategies (Lee et al., 2014). We applied a small volume correction (svc) for the areas of interest with an initial uncorrected threshold of 0.05 on whole-brain-level. The svc was applied on voxel level. Voxels were regarded as significant, when falling below a corrected voxel threshold of 0.05 (family wise error (FWE) corrected) adjusted for the small volume.

2. Control variables

To control for personality traits and behavioral tendencies that may affect flexible learning and decision-making in general, participants filled out several questionnaires at the beginning of the experiment. In particular, participants completed German versions of the State-Trait Anxiety Inventory (STAI; Spielberger et al. 1970), the Trier Inventory of Chronic Stress (TICS; Schulz & Schlotz 1999) and the Beck Depression Inventory (BDI; Beck et al. 1961).

3. Results

3.1. Successful stress induction

Participants first underwent the TSST, a standardized stress protocol consisting of a mock job interview, or a non-stressful control procedure. Subjective and physiological measurements confirmed the successful stress induction through the TSST (Fig. 3A-E). The TSST was experienced as significantly more difficult (t(56) = 5.73, $p = 4.12e^{-07}$, $d = 1.51$), unpleasant (t(56) = 6.70, $p = 1.09\ e^{-08}$, $d = 1.76$), and stressful (t(56) = 5.55, $p = 8.14e^{-07}$, $d = 1.46$) than the control manipulation. Moreover, the TSST, but not the control procedure, led to increased systolic blood pressure (treatment × time: F(3168) = 16.67, p = $1.59e^{-09}$; $\eta^2_{ges} = 0.059$), diastolic blood pressure (F(2.64, 148.01) = 15.67, p = $3.29e^{-08}$ (Greenhouse-Geisser corrected), $\eta^2_{ges} = 0.080$), and pulse (F(2.41, 134.77) = 14.39, p = $3.83e^{-07}$(Greenhouse-Geisser corrected), $\eta^2_{ges} = 0.048$), indicating significant autonomic activation in response to the TSST. Finally, the TSST, but not the control manipulation, induced a pronounced increase in salivary cortisol (treatment × time: F(2.43, 136.31) = 10.70, $p = 3.83e^{-07}$ (Greenhouse-Geisser corrected), $\eta^2_{ges} = 0.0475$). While groups did not differ in cortisol concentrations before the TSST (t(56) = −0.35, p = 0.73, $d = -0.09$), cortisol concentrations were significantly higher in the stress group than in the control group at all time points of measurement after the manipulation (all $p \leq 0.05$). Peak cortisol levels were reached ~18 min after stressor onset, shortly before the Markov decision task in the MRI began, and cortisol levels remained significantly elevated throughout the task.

3.2. Stress reduces the behavioral sensitivity to reversals of reward contingencies

In order to examine how stress changes the flexibility of learning, participants completed a modified Markov decision task in the MRI scanner about 20 min after the onset of the stress or control manipulation. This task was designed to dissociate model-free and model-based learning (Daw et al., 2011; Gläscher et al., 2010) and involved two subsequent choices, each between two fractal stimuli (Fig. 2). The first stage decision led to a second stage, requiring another choice between two options which were associated with different probabilities of monetary reward. Each of the first stage options was predominantly associated with one or the other state in the second stage. Whether or not the transition between the first and the second stage is considered in the decision allows conclusions to be drawn about the underlying learning strategy. While a purely model-free learning strategy only accounts for whether the previous action led to a reward in the second stage, a model-based learner would also include the path that led to the result in the subsequent decision. Learning performance was quantified by the proportion of first stage choices for the stimulus that led predominantly to the second stage state with the overall higher probability to obtain a reward (0.9 | 0.4 vs. 0.6 | 0.1). Likewise, successful learning in the second stage was associated with the proportion of choices for the option with the higher reward probability (either 0.9 or 0.6).

The stress and control groups did not differ in the overall proportion of advantageous choices, neither in the first stage (t(56) = 1.123, $p = 0.266$, $d = 0.295$), nor in the second stage (t(56) = −0.239, $p = 0.81$, $d = -0.062$). This pattern of results is generally in line with previous findings suggesting that the stress-induced alteration in the nature of learning becomes apparent only when the environment changes and the flexibility of behavior is probed (Kim et al., 2001; Schwabe and Wolf, 2009; Schwabe et al., 2010). The proportion of advantageous first stage choices did not differ between blocks (main effect block: F(1, 56) = 0.04), $p = 0.84$, $\eta^2_{ges} = 0.0003$; treatment × block: F(1, 56) = 0.03, $p = 0.87$, $\eta^2_{ges} = 0.0002$), neither did the proportion of advantageous second stage choices (main effect block: F(1, 56) = 1.72), $p = 0.20$, $\eta^2_{ges} = 0.10$; treatment × block: F(1, 56) = 1.23, $p = 0.27$, $\eta^2_{ges} = 0.007$).

In a next step, we analyzed participants' behavioral response to the reversal by comparing the proportion of advantageous choices in the four trials before a reversal relative to the four trials after a reversal between the stress and control groups (Fig. 4). For the first stage choices, the proportion of advantageous choices was – as expected – overall significantly lower after a reversal than before (main effect of time; F(3168) = 25.018, $p = 5.95e^{-06}$, $\eta^2_{ges} = 0.23$), post-hoc *t*-test pre vs. post: t(57) = 4.87, $p = 9.21\ e^{-06}$, $d = 0.64$). Interestingly, the change in first stage choices from pre- to post-reversal differed significantly between groups (treatment × time: F(1, 56) = 4.104, $p = 0.048$, $\eta^2_{ges} = 0.047$). Post-hoc t-tests revealed that the proportion of advantageous choices in the pre-reversal trials was significantly lower in the stress group than in the control group (t(56) = −2.25, $p = 0.03$, $d = -0.59$), while groups did not significantly differ in the proportion of advantageous choices after a reversal (t(56) = 1.19, $p = 0.24$, $d = 0.31$). To verify that the predictions of our model matched the actual data around the reversals, we performed posterior predictive checks. Therefore, we generated 50 simulations for each participant, in which we entered each individual set of optimized parameters into our version of the Markov decision task. Averaging over these simulations we obtained the posterior predictive peri-reversal time course of advantageous choices. This showed a pattern very similar to the actual data (Fig. 4, right panel).

To test whether the differential influence of reversals on choice behavior in the stress group relative to the control group cannot be explained by a general learning impairment in the stress group, we tested whether the proportion of advantageous choice in the stress group differed from chance level. The proportion of advantageous choices in the first stage was significantly different from chance (i.e. 50 percent; t(28) = 3.03, $p = 0.005$), indicating that the stress group had learned the contingencies before a reversal took place. Moreover, the proportion of advantageous choices in the stress group differed in the four trials before a reversal versus four trials after a reversal (t(28) = 2.14, $p = 0.04$, $d = 0.40$), indicating that the reversal did affect the behavior of stressed participants, but to a lesser extent than in controls.

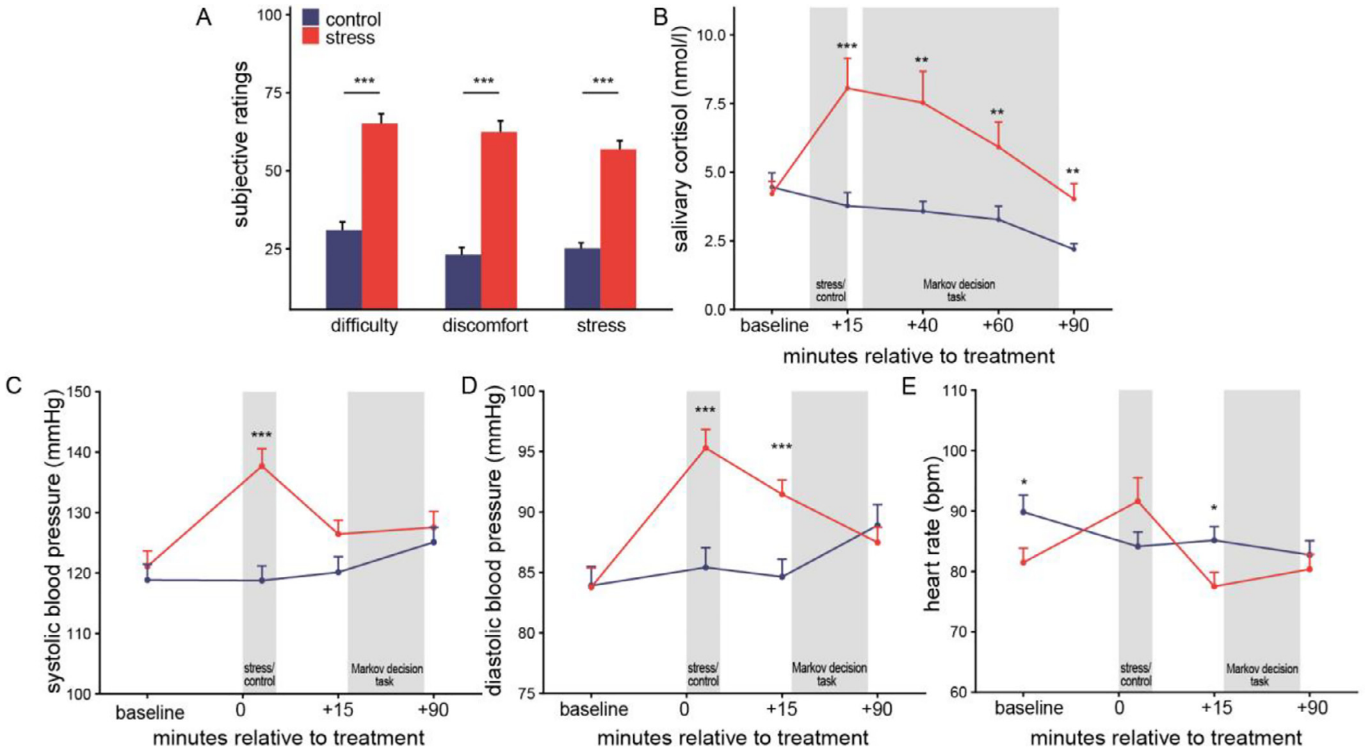


Fig. 3. Successful stress induction. (A) Participants in the stress condition rated the treatment as significantly more difficult, unpleasant, and stressful than participants in the control condition. The exposure to the stressor led further to significant increases in (B) salivary cortisol levels, (C) systolic blood pressure, and (D) diastolic blood pressure. (E) Heart rate measures were significantly lower in the stress group than in the control group at baseline (t(56) = −2.28, $p = 0.03$). Measures increased significantly after stress, relative to baseline (t(28) = 3.34, $p = 0.002$), but decreased at control treatment (t(28) = −4.37, $p = 0.0002$); error bars represent standard errors of the mean, $^{**}p < 0.01$, $^{***}p < 0.001$ for the comparison between the stress group and the control group.

The observed group differences in the first stage are particularly intriguing as the first stage choice indicates the integration of the task structure into the decision. A large proportion of decisions that lead to the better second state suggest an understanding of the state space and the associated transitions (model-based learning) - regardless of the reward obtained in the end. We further tested whether participants' behavior around the reversal, expressed as mean number of advantageous first stage choices in the four trials before a reversal minus the four trials after a reversal, differed between blocks. This analysis showed that neither the stress group (F(1, 28) = 0.04, $p = 0.84$, $\eta^2_{ges} = 0.002$), nor the control group (F(1, 28) = 1.89, $p = 0.18$, $\eta^2_{ges} = 0.06$) changed in their sensitivity to reversals across the three blocks of the experiment.

The proportion of choices for the option with the higher reward probability in the second stage was also significantly lower after a reversal than before (main effect of time; F(1, 56) = 89.948, $p = 3.007e^{-13}$, $\eta^2_{ges} = 0.424$), but did not differ between groups (treatment × time: F(1, 56) = 0.099, $p = 0.755$, $\eta^2_{ges} = 0.0008$). Accordingly, the change in the proportion of advantageous second stage choices from before to after the reversal did not differ between the stress and control groups (t(56) = −0.289, $p = 0.773$, $d = -0.076$; Fig. 4B).

3.3. Model-based and model-free contributions to behavior

In order to capture model-free and model-based contributions to choice behavior, we conducted a logistic regression analysis. The previous trial's transition type and outcome were used to explain whether participants chose the same action again or whether they switched to the other option. This analysis allows a dissociation of model-free and model-based contributions because both learning strategies make qualitatively distinct predictions about how the previous trial's characteristics influence the first stage choice in the following trial. Fig. 5 (left) shows the theory-based choice behavior of purely model-free and model-based learners. A pure model-free strategy predicts that a rewarded action will be repeated, regardless of the transition type (main effect of reward). A model-based agent, on the other hand, uses its knowledge of the task structure and therefore predicts an interaction between transition and reward. The data predicted by our model suggest a mixture of model-free and model-based learning strategies, without differences between the stress group and the control group (Fig. 5, middle).

The logistic regression analysis confirmed the basic signature of model-free reinforcement learning to behavior, indicated by an increased probability to stay when the previous trial was rewarded ($z = 5.715$, $p = 1.10e^{-08,}$ $\beta = 1.295$), as well as the contribution of model-based strategies as indicated by a reward × transition interaction with an additional increase in stay probabilities when a reward was obtained after a common transition ($z = 2.586$, $p = 0.0097$, $\beta = 0.380$). Thus, participants demonstrated both model-based and model-free elements of learning. However, as shown in Fig. 5 (right), the balance of model-based and model-free contributions appeared to be overall biased towards more model-free learning, without significant differences between groups (stress × reward, $z = -1.048$, $p = 0.295$, $\beta = -0.330$; stress × reward × transition, $z = -1.181$, $p = 0.238$, $\beta = -0.235$).

3.4. Stress effects on model-based and model-free parameters

In a next step, we used reinforcement learning models to dissociate model-free and model-based contributions to participants' trial-by-trial choices. We fitted choice behavior to a dual-system reinforcement learning model which includes both model-free and model based learning strategies (Daw et al., 2011; Gläscher et al., 2010). The algorithm contained 7 parameters, fitted individually for each participant.

We assumed that choices were driven by the weighted average of these two computations. The weighting parameter w shows a predominance of model-free proportions in choice behavior (mean control

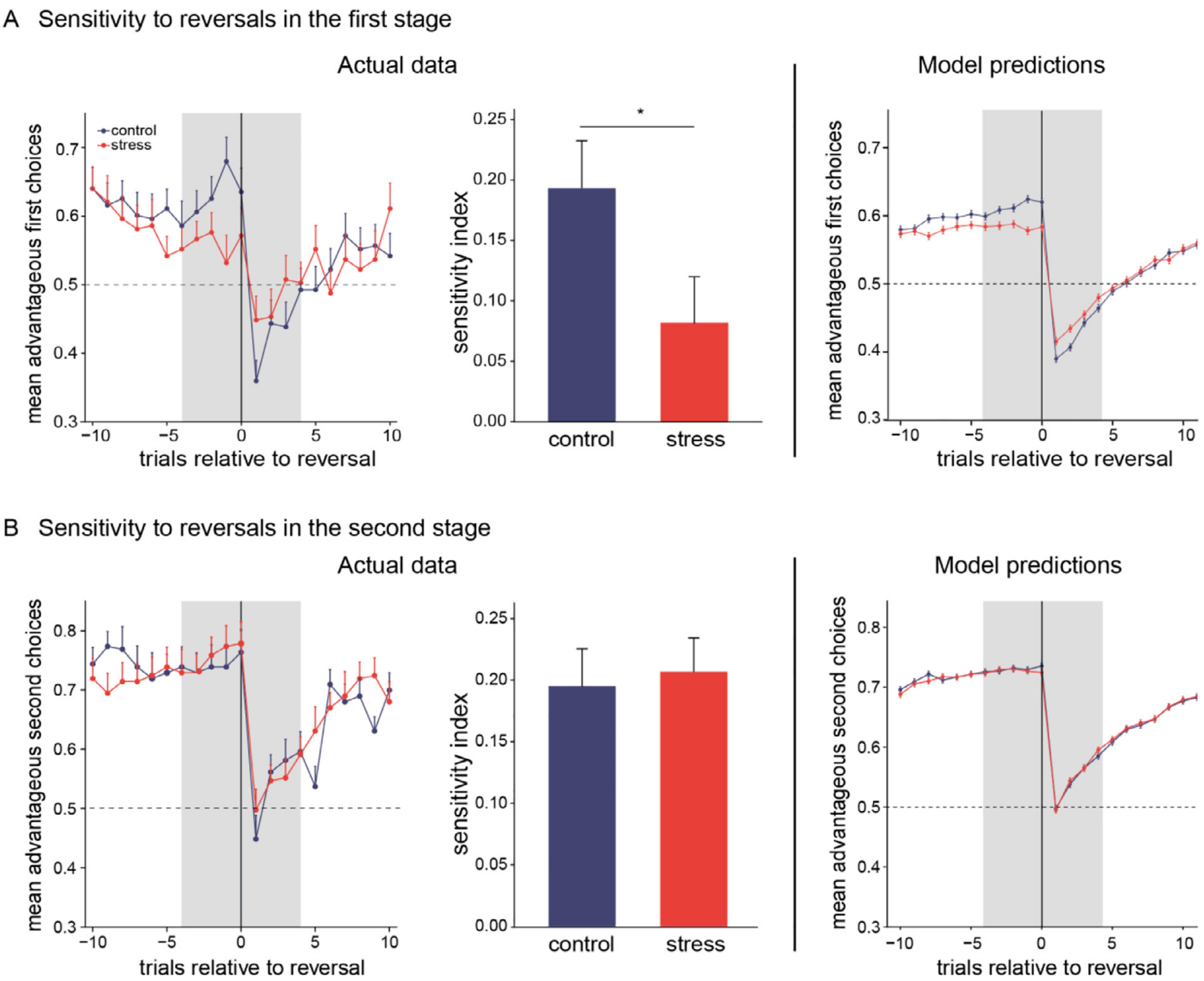


Fig. 4. Stress reduces the behavioral sensitivity to reversals in the first stage. The proportion of advantageous first stage choices is higher in the four trials before a reversal than in the four trials after a reversal, indicating that the reversals have an effect on behavior (A, B). The sensitivity index, computed as the mean of advantageous choices before vs. after a reversal, is significantly higher in the control group than in the stress group in the first stage (A), while the sensitivity index for the second choice does not differ between the stress group and the control group (B). Right panels: Model simulations with best fitting parameters for the trials around the reversals show a pattern similar to the actual behavioral data.

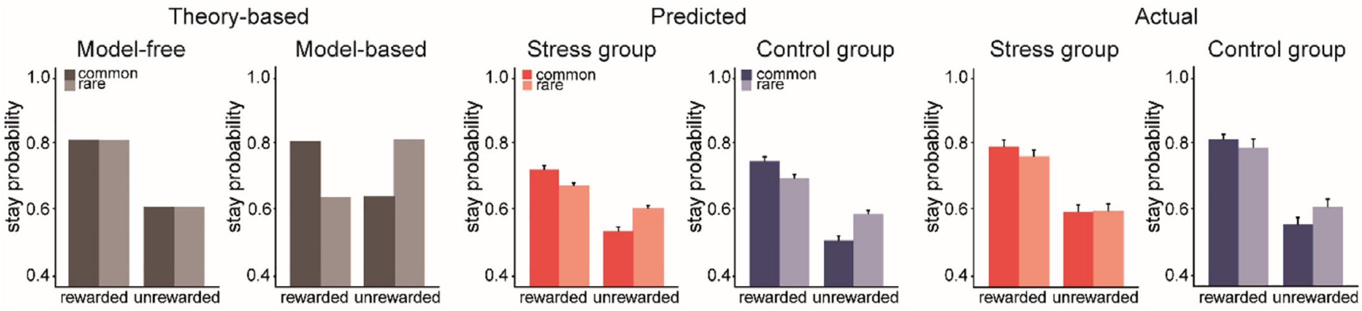


Fig. 5. Factorial analysis of choice behavior. Left: Pure model-free reinforcement learning predicts that a previously rewarded action is more likely to be repeated on the subsequent trial, regardless of whether that reward occurred after a common or a rare transition. Pure model-based behavior comprises a knowledge of the task structure: a reward obtained via a rare transition predicts a switch to the other option. Middle: Data obtained from a posterior predictive check using the set of model parameters estimated for each participant suggests a mixture of both model-free and model-based learning strategies. Right: Actual Data. Participants show both model-based and model-free learning with an overall bias toward model-free learning, independent of group assignment.

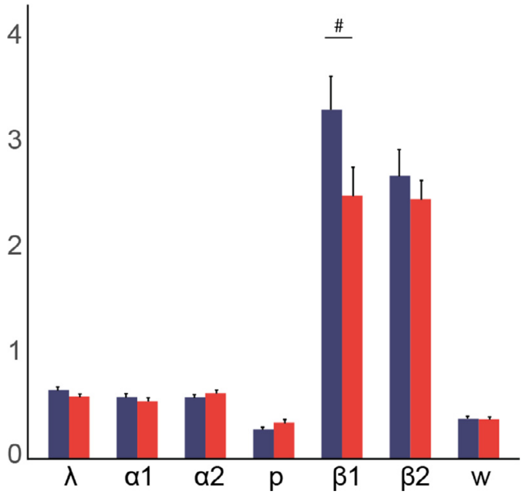


Fig. 6. Stress effects on the model parameters. Best-fitting parameter estimates, shown across subjects. The stress group tended to show a reduced temperature in the first stage, compared to the control group (t(56) = 1.96, $p = 0.056$, $d = 0.51$), indicating more random or exploring choice behavior; no group differences in the eligibility parameter λ, the two learning rates α_1 and α_2, the stay bias p, or the weighting parameter w; error bars represent standard errors of the mean, # $p < 0.06$ for the comparison between the stress group and the control group.

group: 0.38, mean stress group: 0.37, test against 0.50: both $p < 0.001$), which was comparable in the stress and control groups (t(56) = 0.10, $p = 0.918$, $d = 0.03$), indicating that acute stress did not alter the weight of model-free and model-based contributions to learning per se (Fig. 6). However, the temperature parameter for the first stage choice tended to be lower in the stress group (t(56) = 1.96, $p = 0.056$, $d = 0.51$; all other parameters remained unaffected by stress, all $p > 0.20$, Fig. 6). This temperature parameter was significantly positively associated with both the proportion of advantageous first stage choices (r(56) = 0.45, $p = 0.0004$) and the sensitivity index (r(56) = 0.44, $p = 0.0006$). In the second stage, the corresponding temperature parameter was also positively correlated with the proportion of advantageous choices (r(56) = 0.48, $p = 0.0002$), as well as the sensitivity index ($r = 0.48$, $p = 0.0002$). Furthermore, the inverse temperature parameter reflects the extent to which the underlying value computations are used to guide choices, in the sense of an exploration – exploitation trade off or a measure of choice stochasticity. Our results thus point to a rather explorative choice behavior in the stress group, or more random first stage decisions, suggesting that the stressed participants did not use the first decision as a planning step for the second stage, but may have randomly made the first decision in order to proceed to the reward-guided second choice. In other words, stress appeared to affect the utilization of value computations for the first stage choice.

These results are in line with our findings that stress reduced the sensitivity to reversals in the first stage. In addition, we tested whether the sensitivity index correlates with the weighing parameter w. Our results showed no such correlation ($r = -0.17$, $p = 0.19$). The absence of a correlation between the weighing parameter and participants' sensitivity to a reversal was not surprising given that we assume that both model-based and model-free processes may contribute to flexible learning and the sensitivity to changes in the environment. Further modeling parameters did not correlate with the sensitivity index (all r $\langle$ |0.22|, all p $\rangle$ 0.1).

3.5. *Stress affects the neural underpinnings of both model-based and model-free learning*

Our behavioral results suggested that the stress group tended to show more explorative or random choice behavior at the first stage than the control group. Directly building on this pattern of results, we compared the brain activity at advantageous first stage choices with disadvantageous at that time point between the stress and the control group. This analysis showed that stressed participants had significantly reduced activity in the medial prefrontal cortex (mPFC; peak −16 10 62, p_{svc} = 0.03, FWE, Fig. 7B), compared to the control group. Comparing advantageous choices to disadvantageous choices in the second stage, the control group tended to show a higher activity in the ventral striatum (peak 2 10 −8, p_{svc} = 0.07, FWE).

Next, we regressed the model-derived prediction errors against the fMRI data collected during the Markov task. Corroborating earlier reports (Daw et al., 2011; Gläscher et al., 2010; Lee et al., 2014), our data pooled over both groups showed that reward prediction-errors were computed in the lateral OFC, ilPFC, mPFC, ventral striatum, putamen, insula and in the hippocampus (all $p_{svc} < 0.016$, FWE). Reward onsets were associated with activity in the ilPFC, mPFC and insula (all $p_{svc} <$ 0.03, FWE). Interestingly, reward onsets were associated with increased activity in the posterior hippocampus (peak −22 −34 −2, p_{svc} = 0.018, FWE, Fig. 7C) in the stress group. The computation of model-free prediction errors was associated with the lateral OFC, the ventral striatum and the anterior hippocampus (all $p_{svc} < 0.017$, FWE) and model-based prediction errors with activity in the hippocampus, lateral OFC, mPFC and putamen (all $p_{svc} < 0.009$, FWE).

The table shows MNI (Montreal Neurological Institute) coordinates for local maxima in mm. All areas with $k > 5$ significant voxels are reported. For our regions of interest (ROIs), we implemented small volume correction (SVC) using an initial threshold of $p < 0.05$, uncorrected. The significance threshold was set to $p < 0.05$, family wise error (FWE) corrected.

Most interestingly, these neural underpinnings of both model-free and model-based learning were affected by stress (Table 2). Compared to controls, stressed participants showed reduced correlations between BOLD activity and model-free prediction errors in the right ilPFC (peak 48 32 −8, p_{svc} = 0.005, FWE; Fig. 7A) and a tendency to reduced activation in the left amygdala (peak −24 −8 −18, p_{svc} = 0.059, FWE). For model-based prediction errors, stressed participants showed, relative to controls, reduced activity in the right putamen (peak 30 −10 12, p_{svc} = 0.032, FWE, Fig. 7D) and a higher activation in the right ilPFC (peak 48 32 −8, p_{svc} = 0.005, FWE, Fig. 7D). At trend level, stress increased activity in the right insula (peak 32 30 −2, p_{svc} = 0.059, FWE) and led to a decrease in the activity of the right amygdala (peak 30 −4 −20, p_{svc} = 0.054, FWE). Moreover, stress tended to reduce activity in the hippocampus (peak −24 −34 −4, p_{svc} = 0.078, FWE), a region only rather recently implicated in model-based behavior (Vikbladh et al., 2019). Because it is assumed that there is a functional separation along the hippocampal anterior-posterior axis (Fanselow and Dong, 2010; Poppenk et al., 2013; Strange et al., 2014), we further subdivided the hippocampus into anterior and posterior parts, in accordance with previous studies (Collins et al., 2015; Dandolo and Schwabe, 2018), and tested whether the obtained stress effect was specific to the anterior or posterior hippocampus. This analysis revealed that stress affected indeed solely the posterior hippocampal contribution to model-based behavior (peak −24 −34 −2, p_{svc} = 0.019, FWE), while there was no stress effect on the anterior hippocampus (left: p_{svc} = 0.78, right: p_{svc} = 0.17, FWE).

3.6. *Exploratory analysis of control variables and working memory influences*

To control for personality traits and behavioral tendencies that may affect flexible learning or modulate stress effects on flexible learning, we measured state anxiety, trait anxiety, depressive symptoms and chronic stress via the STAI-S, STAI-T, BDI and TICS, respectively. Because one subject code was mistakenly assigned twice, we could not use the questionnaire data of two participants, resulting in n = 56 for the follow-

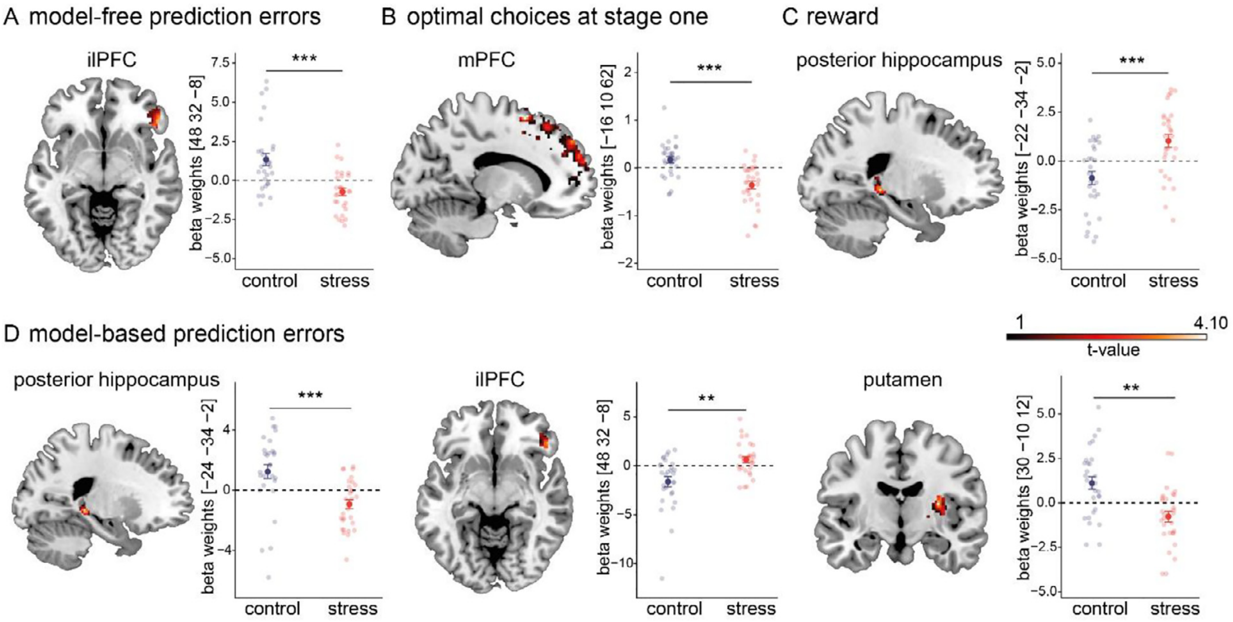


Fig. 7. Stress reduces posterior hippocampal activity during model-based learning and inferiorlateral prefrontal activity during model-free learning. (A) The stress group showed reduced activity during model-free error computation contrary to the control group in the right ilPFC. (B) The stress group showed reduced activity in the mPFC during advantageous choices in the first stage. (C) The stress group showed a higher activity in the posterior hippocampus during reward computations, compared to the control group. (D) Model-based prediction errors were associated with a stress-induced reduction of the posterior hippocampus and the putamen, while the stress group showed an increased activity in the ilPFC, compared to the control group. Data are thresholded at $p < 0.05$, uncorrected, for display purposes only. Parameter estimates were extracted for the peak voxel; error bars represent standard errors of the mean, $^{***}p < 0.001$ for the comparison between the stress group and the control group.

Table 2
Stress effects on neural representations of learning computations.

contrast name	ROI name	Cluster	P_{FWE}	t_{max}	MNI coordinates		
					X	Y	Z
Model-based prediction errors							
control > stress	posterior hippocampus (L)	64	0.019	3.98	−22	−34	−2
control > stress	putamen (R)	74	0.032	4.02	30	−10	12
stress > control	ilPFC (R)	72	0.047	3.61	48	32	−8
Model-free prediction errors							
control > stress	ilPFC (R)	128	0.005	4.58	48	32	−8
Rewards							
stress > control	posterior hippocampus (L)	35	0.018	4.00	−22	−34	−2
Optimal first stage choices							
control > stress	mPFC (L)	26	0.035	4.88	−16	10	62

ing analyses. Importantly, stress and control groups did not differ in depressive symptoms (t(54) = 1.62, p = 0.11, d = 0.43), state anxiety (t(54) = 0.33, p = 0.74, d = 0.089), trait anxiety (t(54) = 1.16, p = 0.25, d = 0.31), or subjective chronic stress (t(54) = 0.89, p = 0.38, d = 0.24, Table 1). Furthermore, in light of previous evidence suggesting that anxiety, depressive symptoms or chronic stress may be associated with the vulnerability to stress and changes in model-based behavior (Nasca et al., 2015; Radenbach et al., 2015; Weger and Sandi, 2018), we further tested whether the questionnaire data correlated with the sensitivity index or model-derived parameters. These analyses yielded no significant correlations between the sensitivity index and state / trait anxiety, chronic stress, or depressive symptoms (stress: all |r| < 0.16, all p > 0.4, control: all |r| ⟨ 0.37, all p ⟩ 0.06, all participants: all |r| ⟨ 0.25, all p ⟩ 0.06), except for a significant negative correlation between STAI-S scores and the sensitivity index in the control group (r = −0.418, p = 0.03), which would however not survive a correction for multiple comparisons. When we subdivided participants into subgroups based on a median-split on the respective questionnaire score, we obtained evidence suggesting that acute stress might influence participants' behavioral response to the reversal in particular in individuals with high trait or state anxiety. Further, stress and control groups appeared to differ in particular when participants reported low chronic stress and low levels of depressive mood (see supplemental Figure S1 and supplemental Table S2). These analyses, however, were exploratory and need to be interpreted with great caution.

Because there is evidence that high baseline working memory might protect model-based learning from deleterious stress effects (Otto et al., 2013), participants completed an n-back test, as common measure of working memory (Owen et al., 2005), before they underwent the stress or control manipulation. The working memory data of four participants are missing due to technical failure. Importantly, groups did not differ in baseline working memory performance (t(52) = −1.38, p = 0.17, d = −0.38). When we analyzed correlations between baseline working memory performance on the one hand and the task performance (i.e. the sensitivity index) on the other hand, we obtained no significant correlations, neither within the stress or control groups (stress: r(23) = 0.293, p = 0.155; control: r(27) = −0.004, p = 0.984), nor across all participants (r = 0.178, p = 0.197). These correlational data suggest that base-

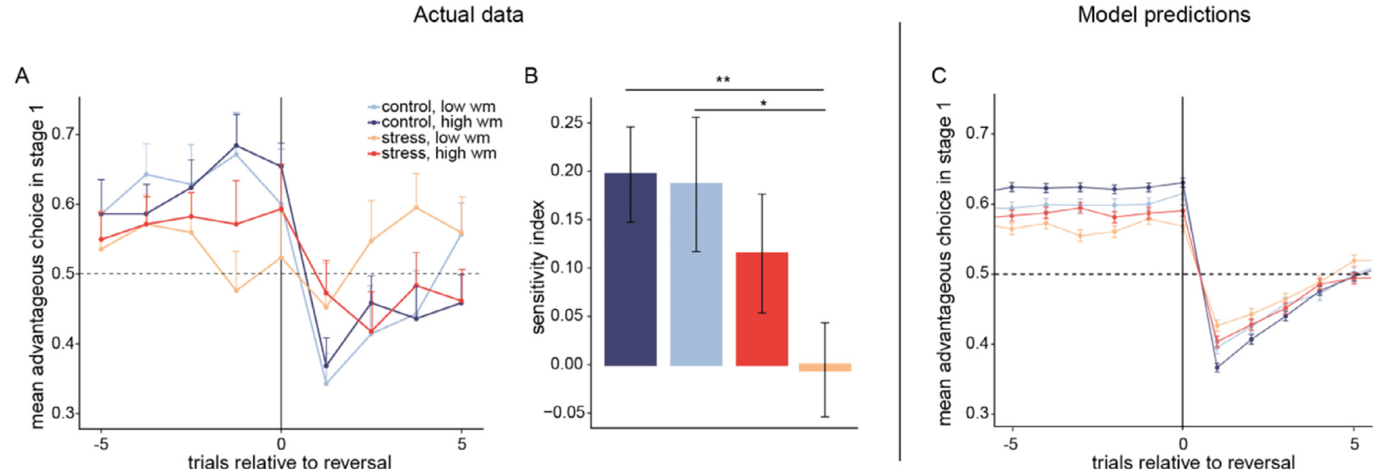


Fig. 8. Stress effects, separately for high and low baseline working memory capacity, as measured with an n-back task. (A) Our data suggest that subjects with low working memory are particularly susceptible to stress effects on flexible learning, yet the interaction between stress and working memory is not statistically significant. (B) The sensitivity index, computed by the mean of advantageous choices before vs. after a reversal, is significantly higher in the control group (high and low working memory) than in the low working memory stress group (t(39) = 2.88, p = 0.006. d = 0.99). (C) Posterior predictive behavior in the trials around the reversal, separately for individuals with high and low working memory capacity in the stress group and in the control group, confirms that the model predictions match the actual behavior, except for a deviation in the stress/low working memory group; $^{***}p$ < 0.001, $^{**}p$ < 0.01 and $^{*}p$ < 0.05 for the comparison between the groups; error bars represent standard errors of the mean.

line working memory does not modulate the impact of stress on learning performance. However, it may also be assumed that a differential susceptibility to stress effects is less modulated by gradual differences in the working memory, but is rather apparent at particularly high or particularly low scores of working memory. Therefore, we tested in a next step whether stress affected the proportion of advantageous first stage choices 4 trials before vs. 4 trials after a reversal differently in high vs. low working memory groups, respectively. High and low working memory groups were defined based on a median split on the n-back performance. The performance of the high- and low working memory participants in the stress and control groups is shown in Fig. 8A. Although Figs. 8A and B suggest that the sensitivity for the change in reward contingencies was particularly affected in stressed participants with low baseline working memory capacity (t(39) = 2.88, p = 0.0065), while the high-working memory stress group and the control group did not differ (t(40) = 1.09, p = 0.282), the respective working memory × stress interaction was not statistically significant (F(52) = 0.89, p = 0.35, η^2_{ges} = 0.01).

Again, we tested whether our model's predictions matched the pattern found in the behavioral data around the reversals and therefore generated 50 simulations for each participant's individual set of parameters. These simulations showed a pattern that strongly resembled the actual data, except for the stress/low working memory group (Fig. 8C). For this group, the correspondence between the simulated and the actual data was lower. In the actual data, the behavior is hardly influenced by the contingency changes, while the simulations show a decrease of the advantageous decisions after a reversal. However, the order of the four groups in the posterior predictive behavior is broadly consistent with the measured data, i.e. also in the simulated data the stress/low working memory group shows the smallest difference from pre- to post-reversal. However, the difference between pre- and post-reversal can still be clearly seen in the simulations, which is not reflected in the actual data. This can be explained by the much smaller sub-sample size in the measured data (data n = {10,12,13,19} vs. 50 in the simulations). On the other hand, this indicates that there are other sources of noise in the measured data that cannot be mapped with the learning model of the Markov decision task.

After analyzing the influence of working memory capacity on the performance in the Markov decision task, we investigated whether working memory was associated with the model parameters. Scores in the n-back task were overall positively correlated with the temperature parameter in stage 1 (r = 0.3, p = 0.03) and tended to be associated with a higher temperature parameter in the second stage (r = 0.26, p = 0.06). Given that the temperature parameter determines to which extent the learned information is used to guide subsequent choices, the observed link to working memory processes is not surprising and might also point to general cognitive capacities that contribute to both flexible learning and working memory. Moreover, high n- back scores tended to be associated with a lower learning rate in the second stage (r = −0.26, p = 0.06, supplemental table S3). However, there were no significant correlations between working memory and model parameters in the stress and control groups (all p > 0.177) and there were no significant main or interaction effects including the factor stress in our working memory × group ANOVA (all F(50) ⟨ 1.78, all p ⟩ 0.19).

4. Discussion

Successful adaptation to dynamic environments is crucial for survival, particularly under highly stressful or threatening conditions. Stress, however, is assumed to impede behavioral flexibility (Otto et al., 2013; Plessow et al., 2011; Raio et al., 2017; Schwabe and Wolf, 2011; Vogel et al., 2016). Here, we sought to shed light on the neurocomputational mechanisms involved in the stress-induced deficit in flexible learning. Our behavioral data show that stress indeed reduced participants' sensitivity to changes in outcome contingencies. In line with these data, our model-based analyses suggest that stress tended to favor rather explorative behavior, as reflected in the tendency of a reduced softmax temperature for the first stage decision. We assume that this is moderated by a reduced utilization of value signals negotiated by model-based and model-free processes. Most importantly, our model-based fMRI analyses revealed that stress reduced the contributions of structures implicated in model-based control and those involved in model-free control of learning.

To tackle specifically the flexibility of learning, we modified the original Markov decision task (Daw et al., 2011) by including several reversals in reward contingencies. This modification increased the task difficulty and made it more demanding to establish a valid model of the task structure, thus favoring, irrespective of stress, model-free over model-based learning. Indeed, although we obtained clear evidence for model-based contributions, model-free elements prevailed during learning. This overall bias towards more model-free learning, reflected in participants' stay probabilities and the weighing parameter w, corroborates recent research suggesting that task complexity facilitates an increased reliance on model-free learning (Kim et al., 2018). The task-related bias towards more model-free learning may explain why we did not observe a further stress-induced shift towards model-free learning that has been suggested before (Park et al., 2017). Accordingly, the proposed bias towards model-free learning associated with the modified task might be considered a limitation, although it is to be noted that participants' choice behavior and the computational modeling parameters provided evidence for both model-based and model-free learning mechanisms. Our behavioral data point to an impairment of flexible learning that is not owing to an altered balance of model-based and model-free processes but rather to a reduced contribution of both model-based and model-free processes to behavior, in contrast to earlier findings suggesting mainly a stress-induced impairment of model-based learning (Otto et al., 2013). The observed impairment seemed to be most pronounced in individuals with low working memory capacity. Although the respective interaction effect did not reach statistical significance, this pattern is generally in line with evidence suggesting that a high working memory capacity may prevent stress effects on model-based learning (Otto et al., 2013).

Although our behavioral data may be interpreted as an indication of impaired flexible learning after stress due to reduced sensitivity for reversals, an alternative view would be that stress encourages more explorative choices at the first stage. More specifically, stressed participants may have learned the stimulus-action-reward-associations in the same way as controls but nevertheless tend not to use this information to guide their behavior. This is indicated by the trend towards a stress-induced reduction of the first stage temperature parameter and further supported by the positive correlation between the first stage sensitivity index and both stage's temperature parameters. At first glance, these findings might seem to be in conflict with previous findings suggesting that stress leads to rather exploitative decisions (Lenow et al., 2017; Luksys and Sandi, 2011). However, these previous studies used classical foraging tasks and such tasks require a different type of decision-making in which the overall environment is used as a proxy for the value of future unknown options, compared to current prospects. Thereby, the focus is on reward calculations which usually determine the switch to a new option below a certain threshold, while the focus in the present task is to maintain probabilistic rules to guide actions. Therefore, another possible explanation is that working memory mediates the explorative choice behavior in the first stage, given that exploration could also be due to an inability to maintain the relevant information to guide upcoming decisions. In line with this idea, performance appeared to be particularly explorative after stress in participants with low working memory performance. Increased explorative behavior in this task can be both advantageous and disadvantageous: it prevents the reliable repetition (exploitation) of a learned contingency but protects against a performance drop when contingencies change. This could explain why the proportion of advantageous decisions did not differ between the groups overall, while there were group differences in the trials around the reversals.

Our data provided initial evidence that stressed participants use value information less for their decision in the first, but not the second stage, as indicated by the softmax temperature parameter and the sensitivity index. This view is further supported by a significantly reduced sensitivity index in the stressed participants with low working memory capacity, given the fact that working memory holds behaviorally relevant information to guide action. The stress-related impairment in first stage choices was accompanied by reduced activity in the ilPFC in the stress group during first stage onset compared to the control group. Thus, the reduced behavioral sensitivity to reversals may be owing to detrimental stress effects on the ilPFC, which has previously been linked to the arbitration between model-based and model-free learning (Lee et al., 2014).

In support of the view that stress interfered with both model-based and model-free control, our imaging findings showed that stress affected the neural underpinnings of both model-free and model-based learning. More specifically, stress reduced the activity associated with model-free prediction errors in the ilPFC. At the same time, the stress group showed an increase in ilPFC activity during model-based learning. The ilPFC has been associated with an arbitrator signal that determines whether behavior is guided by model-based or model-free learning systems (Lee et al., 2014). It is assumed that this arbitrator reduces activity in brain areas implicated in model-free learning when the arbitrator deems that behavior should be guided by the model-based system (Lee et al., 2014). Accordingly, a stress-induced increase in ilPFC activity related to model-based learning processes may be paralleled by a decrease or suppression of the model-free system, as observed here. At the same time, stress decreased activity during model-based prediction error computations in the putamen and posterior hippocampus. In particular the hippocampus has very recently been implicated in model-based planning (Schuck and Niv, 2019; Vikbladh et al., 2019). In fact, the idea of a cognitive map stored in the hippocampus has been proposed already several decades ago (O'Kneefe and Nadel, 1978). For long, however, this idea was limited to spatial references. A recent integrative approach suggests that the hippocampus may also encode cognitive maps that capture complex relationships between cues, actions, results and other characteristics of the environment, enabling flexible, goal-directed decision making (Wikenheiser and Schoenbaum, 2016). Importantly, however, the hippocampus may not as whole be involved in model-based learning. Accumulating evidence from human neuroimaging and rodent lesion studies suggests a functional dissociation within the hippocampus along its anterior (ventral) – posterior (dorsal) axis (Poppenk et al., 2013; Zeidman and Maguire, 2016). In a recent rodent study, specifically the dorsal (posterior) hippocampus was linked to model-based planning behavior (Miller et al., 2017). This finding dovetails with the present data showing that stress reduced specifically the posterior hippocampal activity associated with model-based prediction errors.

These neural changes are most-likely driven by the many hormones and neurotransmitters that are released in response to stressful encounters. Receptors for these stress mediators are abundantly expressed in those regions involved in model-based and model-free learning, in particular, in prefrontal and limbic areas (Herman et al., 2003). Accordingly, it has been shown across tasks and species that these stress mediators, including catecholamines and glucocorticoids, may affect prefrontal and limbic activity and function (Arnsten, 2009; J. J. Kim and Diamond, 2002). Most interestingly with respect to the present findings, it has been shown that glucocorticoids may reduce specifically posterior medial temporal activity during a declarative memory task (de Quervain et al., 2003), exactly that region that was reduced by stress during model-based processing.

Finally, one might argue that the modification of the original Markov decision task impacts the assessment of model-based and model-free processes in our task. While the present task modification, which was required to probe flexible learning in a highly volatile environment, might complicate the direct comparison to studies using the classical Markov decision task to some extent, we assume that also the modified task version allows the assessment of model-based and model-free processes. First, participants' choice behavior and our modeling parameters provided evidence for the involvement of model-based and model-free learning mechanisms. Furthermore, our neuroimaging data revealed neural activity patterns that are well in line with the previously reported neural signatures of model-based and model-free learning, respectively (Daw et al., 2011; Gläscher et al., 2010; Vikbladh et al., 2019). Moreover, the contingency reversals required participants to learn the new transitions, whereas these transitions were assumed to be known by our model. To test whether this affected the performance of our model, we implemented an enhanced model that included state space learning as used in Gläscher et al. (2010) (see Methods for details). Model simulations showed only slightly worse model performance that this enhanced model and our winning model were very similar in their capacity to fit the experimental data. Further analyses also revealed that the transition probabilities in the Markov decision task are learned within the first 10 trials. This supports our original model choice, in which state space learning was omitted.

Together, our data show that stress reduces both model-free and model-based computations during learning in a highly volatile environment. These findings provide novel insights into the neurocomputational mechanisms through which stress hampers the cognitive adaptation to highly volatile environments. A better understanding of these mechanisms may aid the development of new approaches to prevent such stress-induced deficits, with considerable implications, for instance, for educational settings (Vogel and Schwabe, 2016) and stress-related psychopathologies characterized by a deficit in the flexible adaptation to dynamic environments (Koob and Kreek 2007; LaGarde et al., 2010; de Quervain et al. 2017).

Author contributions

L.S. and J.G. conceived and designed the experiment, A.C. performed research, A.C. and F.K. analyzed the data, J.G. fit the computational models, L.S. and J.G. supervised research and analysis, A.C. and L.S. drafted the manuscript, all authors contributed to the manuscript.

Data availability statement

The data that support the findings of this study are openly available on OSF at https://osf.io/tm7ez/?view_only=9d620c297db8461283f01b17a3fa4574.

Declaration of Competing Interest

The authors declare no competing financial interests.

Acknowledgments

This work was supported by the Landesforschungsförderung Hamburg (LFF FV38). We thank Anna Neuhäußer, Theresa Ruder, Leonie el Khafif, Lucie Lilienthal and Rosann Stocker for their assistance with data collection and Carlo Hiller for programming the task.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2021.117747.

References

Aarts, E., Roelofs, A., van Turennout, M., 2008. Anticipatory activity in anterior cingulate cortex can be independent of conflict and error likelihood. J. Neurosci. 28 (18), 4671–4678. doi:10.1523/JNEUROSCI.4400-07.2008.

Alexander, W.H., Brown, J.W., 2011. Medial prefrontal cortex as an action-outcome predictor. Nat. Neurosci. 14 (10), 1338–1344. doi:10.1038/nn.2921.

Arnsten, A.F.T, 2009. Stress signalling pathways that impair prefrontal cortex structure and function. Nat. Rev. Neurosci. 10 (6), 410–422. doi:10.1038/nrn2648.

Balleine, B.W., O'doherty, J.P., 2010. Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. Neuropsychopharmacology 35 (1), 48.

Bayer, H.M., Glimcher, P.W., 2005. Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 47 (1), 129–141. doi:10.1016/j.neuron.2005.05.020.

Beck, A.T., Ward, C., Mendelson, M., Mock, J., Erbaugh, J., 1961. Beck depression inventory (BDI). Arch. Gen. Psychiatry 4 (6), 561–571.

Braun, S., Hauber, W., 2013. Acute stressor effects on goal-directed action in rats. Learn. Mem. 20 (12), 700–709. doi:10.1101/lm.032987.113.

Brown, J.W., 2009. Conflict effects without conflict in anterior cingulate cortex: multiple response effects and context specific representations. Neuroimage 47 (1), 334–341. doi:10.1016/j.neuroimage.2009.04.034.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A., 2017. Stan: a probabilistic programming language. J. Stat. Softw. 76 (1). doi:10.18637/jss.v076.i01.

Collin, S.H.P., Milivojevic, B., Doeller, C.F, 2015. Memory hierarchies map onto the hippocampal long axis in humans. Nat. Neurosci. 18 (11), 1562–1564. doi:10.1038/nn.4138.

Croxson, P.L., Walton, M.E., O'Reilly, J.X., Behrens, T.E.J., Rushworth, M.F.S, 2009. Effort-based cost-benefit valuation and the human brain. J. Neurosci. 29 (14), 4531–4541. doi:10.1523/JNEUROSCI.4515-08.2009.

Dandolo, L.C., Schwabe, L., 2018. Time-dependent memory transformation along the hippocampal anterior–posterior axis. Nat. Commun. 9 (1). doi:10.1038/s41467-018-03661-7.

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. doi: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. Nat. Neurosci. 8 (12), 1704–1711. doi:10.1038/nn1560.

de Quervain, D., Henke, K., Aerni, A., Treyer, V., McGaugh, J.L., Berthold, T., Nitsch, R.M., Buck, A., Roozendaal, B., Hock, C., 2003. Glucocorticoid-induced impairment of declarative memory retrieval is associated with reduced blood flow in the medial temporal lobe: glucocorticoids impair memory retrieval: a PET-study. Eur. J. Neurosci. 17 (6), 1296–1302. doi:10.1046/j.1460-9568.2003.02542.x.

de Quervain, D., Schwabe, L., Roozendaal, B., 2017. Stress, glucocorticoids and memory: implications for treating fear-related disorders. Nat. Rev. Neurosci. 18 (1), 7–19. doi:10.1038/nrn.2016.155.

Deichmann, R., Gottfried, J.A., Hutton, C., Turner, R., 2003. Optimized EPI for fMRI studies of the orbitofrontal cortex. Neuroimage 19 (2), 430–441. doi:10.1016/S1053-8119(03)00073-9.

Diamond, D.M., Campbell, A.M., Park, C.R., Halonen, J., Zoladz, P.R., 2007. The temporal dynamics model of emotional memory processing: a synthesis on the neurobiological basis of stress-induced amnesia, flashbulb and traumatic memories, and the Yerkes-Dodson Law. Neural Plast. 2007, 1–33. doi:10.1155/2007/60803.

Dolan, R.J., Dayan, P., 2013. Goals and habits in the brain. Neuron 80 (2), 312–325. doi:10.1016/j.neuron.2013.09.007.

Fanselow, M.S., Dong, H.-.W., 2010. Are the dorsal and ventral hippocampus functionally distinct structures? Neuron 65 (1), 7–19. doi:10.1016/j.neuron.2009.11.031.

Faul, F., Erdfelder, E., Buchner, A., Lang, A.-.G, 2007. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav. Res. Methods 39, 175–191.

Garvert, M.M., Dolan, R.J., Behrens, T.E., 2017. A Map of Abstract Relational Knowledge in the Human Hippocampal–Entorhinal Cortex, 6. ELife doi:10.7554/eLife.17086.

Gershman, S.J., Uchida, N., 2019. Believing in dopamine. Nat. Rev. Neurosci. 20 (11), 703–714. doi:10.1038/s41583-019-0220-7.

Gläscher, J., Daw, N., Dayan, P., O'Doherty, J.P., 2010. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron 66 (4), 585–595. doi:10.1016/j.neuron.2010.04.016.

Goldfarb, E.V., Sinha, R., 2018. Drug-induced glucocorticoids and memory for substance use. Trends Neurosci. 41 (11), 853–868. doi:10.1016/j.tins.2018.08.005.

Goodman, J., Leong, K.-.C., Packard, M.G., 2012. Emotional modulation of multiple memory systems: implications for the neurobiology of post-traumatic stress disorder. Rev. Neurosci. 23, 5–6. doi:10.1515/revneuro-2012-0049.

Gourley, S.L., Swanson, A.M., Jacobs, A.M., Howell, J.L., Mo, M., DiLeone, R.J., Koleske, A.J., Taylor, J.R., 2012. Action control is mediated by prefrontal BDNF and glucocorticoid receptor binding. Proc. Natl. Acad. Sci. 109 (50), 20714–20719. doi:10.1073/pnas.1208342109.

Haruno, M., Kawato, M., 2006. Different neural correlates of reward expectation and reward expectation error in the putamen and caudate nucleus during stimulus-action-reward association learning. J. Neurophysiol. 95 (2), 948–959. doi:10.1152/jn.00382.2005.

Herman, J.P., Figueiredo, H., Mueller, N.K., Ulrich-Lai, Y., Ostrander, M.M., Choi, D.C., Cullinan, W.E., 2003. Central mechanisms of stress integration: hierarchical circuitry controlling hypothalamo–pituitary–adrenocortical responsiveness. Front. Neuroendocrinol. 24 (3), 151–180. doi:10.1016/j.yfrne.2003.07.001.

Kim, D., Park, G.Y., O'Doherty, J.P., Lee, S.W, 2018. Task complexity interacts with state-space uncertainty in the arbitration process between model-based and model-free reinforcement-learning at both behavioral and neural levels. BioRxiv doi:10.1101/393983.

Kim, J.J., Diamond, D.M., 2002. The stressed hippocampus, synaptic plasticity and lost memories. Nat. Rev. Neurosci. 3 (6), 453–462. doi:10.1038/nrn849.

Kim, J.J., Lee, H.J., Han, J.-.S., Packard, M.G., 2001. Amygdala is critical for stress-induced modulation of hippocampal long-term potentiation and learning. J. Neurosci. 21 (14), 5222–5228. doi:10.1523/JNEUROSCI.21-14-05222.2001.

Kirchner, W.K., 1958. Age differences in short-term retention of rapidly changing information. J. Exp. Psychol. 55 (4), 352–358. doi:10.1037/h0043688.

Kirschbaum, C., Kudielka, B.M., Gaab, J., Schommer, N.C., & Hellhammer, D.H. (1999). Impact of gender, menstrual cycle phase, and oral contraceptives on the activity of the hypothalamus-pituitary-adrenal axis: *psychosomatic medicine*, 61(2), 154–162. 10.1097/00006842-199903000-00006

Kirschbaum, C., Pirke, K.-.M., Hellhammer, D.H., 1993. The 'trier social stress test' – a tool for investigating psychobiological stress responses in a laboratory setting. Neuropsychobiology 28 (1–2), 76–81. doi:10.1159/000119004.

Koob, G., Kreek, M.J., 2007. Stress, dysregulation of drug reward pathways, and the transition to drug dependence. Am. J. Psychiatry 164 (8), 1149–1159. doi:10.1176/appi.ajp.2007.05030503.

LaGarde, G., Doyon, J., Brunet, A., 2010. Memory and executive dysfunctions associated with acute posttraumatic stress disorder. Psychiatry Res. 177 (1–2), 144–149. doi:10.1016/j.psychres.2009.02.002.

Lancaster, J.L., Woldorff, M.G., Parsons, L.M., Liotti, M., Freitas, C.S., Rainey, L., Kochunov, P.V., Nickerson, D., Mikiten, S.A., Fox, P.T., 2000. Automated Talairach Atlas labels for functional brain mapping. Hum. Brain Mapp. 10 (3), 120–131 10.1002/1097-0193(200007)10:3<120::AID–HBM30>3.0.CO;2-8.

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. doi:10.1016/j.neuron.2013.11.028.

Lenow, J.K., Constantino, S.M., Daw, N.D., Phelps, E.A., 2017. Chronic and acute stress promote overexploitation in serial decision making. J. Neurosci. 37 (23), 5681–5689. doi:10.1523/JNEUROSCI.3618-16.2017.

Luksys, G., Sandi, C., 2011. Neural mechanisms and computations underlying stress effects on learning and memory. Curr. Opin. Neurobiol. 21 (3), 502–508. doi:10.1016/j.conb.2011.03.003.

Lupien, S.J., McEwen, B.S., Gunnar, M.R., Heim, C., 2009. Effects of stress throughout the lifespan on the brain, behaviour and cognition. Nat. Rev. Neurosci. 10 (6), 434–445. doi:10.1038/nrn2639.

Maldjian, J.A., Laurienti, P.J., Kraft, R.A., Burdette, J.H., 2003. An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. Neuroimage 19 (3), 1233–1239. doi:10.1016/S1053-8119(03)00169-1.

McClure, S.M., Berns, G.S., Montague, P.R., 2003. Temporal prediction errors in a passive learning task activate human striatum. Neuron 38 (2), 339–346. doi:10.1016/S0896-6273(03)00154-5.

Miller, K.J., Botvinick, M.M., Brody, C.D., 2017. Dorsal hippocampus contributes to model-based planning. Nat. Neurosci. 20 (9), 1269–1276. doi:10.1038/nn.4613.

Nasca, C., Bigio, B., Zelli, D., Nicoletti, F., McEwen, B.S., 2015. Mind the gap: glucocorticoids modulate hippocampal glutamate tone underlying individual differences in stress susceptibility. Mol. Psychiatry 20 (6), 755–763. doi:10.1038/mp.2014.96.

Nee, D.E., Kastner, S., Brown, J.W., 2011. Functional heterogeneity of conflict, error, task-switching, and unexpectedness effects within medial prefrontal cortex. Neuroimage 54 (1), 528–540. doi:10.1016/j.neuroimage.2010.08.027.

O'Doherty, J.P., Dayan, P., Friston, K., Critchley, H., Dolan, R.J., 2003. Temporal difference models and reward-related learning in the human brain. Neuron 38 (2), 329–337.

O'Kneefe, J., Nadel, L., 1978. The Hippocampus as a Cognitive Map. Oxford University Press.

Oliveira, F.T.P., McDonald, J.J., & Goodman, D. (2007). *Performance monitoring in the anterior cingulate is not all error related: expectancy deviation and the representation of action–outcome associations*. 19(12), 11.

Otto, A.R., Raio, C.M., Chiang, A., Phelps, E.A., Daw, N.D., 2013. Working-memory capacity protects model-based learning from stress. Proc. Natl. Acad. Sci. 110 (52), 20941–20946. doi:10.1073/pnas.1312011110.

Owen, A.M., McMillan, K.M., Laird, A.R., Bullmore, E., 2005. N-back working memory paradigm: a meta-analysis of normative functional neuroimaging studies. Hum. Brain Mapp. 25 (1), 46–59. doi:10.1002/hbm.20131.

Park, H., Lee, D., Chey, J., 2017. Stress enhances model-free reinforcement learning only after negative outcome. PLoS ONE 12 (7), 1–12.

Pfeiffer, B.E., Foster, D.J., 2013. Hippocampal place-cell sequences depict future paths to remembered goals. Nature 497 (7447), 74–79. doi:10.1038/nature12112.

Pinheiro, J.C., Bates, D.M., 2000. Mixed-Effects Models in S and S-PLUS. Springer.

Plessow, F., Fischer, R., Kirschbaum, C., Goschke, T., 2011. Inflexibly focused under stress: acute psychosocial stress increases shielding of action goals at the expense of reduced cognitive flexibility with increasing time lag to the stressor. J. Cogn. Neurosci. 23 (11), 3218–3227.

Poppenk, J., Evensmoen, H.R., Moscovitch, M., Nadel, L., 2013. Long-axis specialization of the human hippocampus. Trends Cogn. Sci. 17 (5), 230–240. doi:10.1016/j.tics.2013.03.005.

R Core Team, 2019. R: A language and Environment for Statistical Computing. R Foundation for Statistical Computing https://www.R-project.org/.

Radenbach, C., Reiter, A.M.F., 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. doi:10.1016/j.psyneuen.2014.12.017.

Raio, C.M., Hartley, C.A., Orederu, T.A., Li, J., Phelps, E.A., 2017. Stress attenuates the flexible updating of aversive value. Proc. Natl. Acad. Sci. 201702565.

Rohleder, N., Kirschbaum, C., 2006. The hypothalamic–pituitary–adrenal (HPA) axis in habitual smokers. Int. J. Psychophysiol. 59 (3), 236–243. doi:10.1016/j.ijpsycho.2005.10.012.

Roozendaal, B., McEwen, B.S., Chattarji, S., 2009. Stress, memory and the amygdala. Nat. Rev. Neurosci. 10 (6), 423–433. doi:10.1038/nrn2651.

Schuck, N.W., Niv, Y., 2019. Sequential replay of nonspatial task states in the human hippocampus. Science 364 (6447). doi:10.1126/science.aaw5181, eaaw5181.

Schulz, P., Schlotz, W., 1999. Trierer Inventar zur Erfassung von chronischem Sre (TICS): skalenkonstruktion, teststatistische Überprüfung und Validierung der Skala Arbeitsüberlastung. [The Trier Inventory for the Assessment of Chronic Stress (TICS). Scale construction, statistical testing, and validation of the scale work overload.]. Diagnostica 45 (1), 8–19. doi:10.1026//0012-1924.45.1.8.

Schwabe, L., Joëls, M., Roozendaal, B., Wolf, O.T., Oitzl, M.S., 2012a. Stress effects on memory: an update and integration. Neurosci. Biobehav. Rev. 36 (7), 1740–1749. doi:10.1016/j.neubiorev.2011.07.002.

Schwabe, L., Tegenthoff, M., Hoffken, O., Wolf, O.T., 2012b. Simultaneous glucocorticoid and noradrenergic activity disrupts the neural basis of goal-directed action in the human brain. J. Neurosci. 32 (30), 10146–10155. doi:10.1523/JNEUROSCI.1304-12.2012.

Schwabe, L., Wolf, O.T., 2009. Stress prompts habit behavior in humans. J. Neurosci. 29 (22), 7191–7198. doi:10.1523/JNEUROSCI.0979-09.2009.

Schwabe, L., Wolf, O.T., 2012. Stress modulates the engagement of multiple memory systems in classification learning. J. Neurosci. 32 (32), 11042–11049. doi:10.1523/JNEUROSCI.1484-12.2012.

Schwabe, L., Schächinger, H., de Kloet, E.R., Oitzl, M.S., 2010. Corticosteroids operate as a switch between memory systems. J. Cogn. Neurosci. 22 (7), 1362–1372. doi:10.1162/jocn.2009.21278.

Schwabe, L., Tegenthoff, M., Höffken, O., Wolf, O.T., 2013. Mineralocorticoid receptor blockade prevents stress-induced modulation of multiple memory systems in the human brain. Biol. Psychiatry 74 (11), 801–808. doi:10.1016/j.biopsych.2013.06.001.

Schwabe, L., Wolf, O.T., 2011. Stress-induced modulation of instrumental behavior: from goal-directed to habitual control of action. Behav. Brain Res. 219 (2), 321–328. doi:10.1016/j.bbr.2010.12.038.

Schwabe, L., Wolf, O.T., 2013. Stress and multiple memory systems: from 'thinking' to 'doing.'. Trends Cogn. Sci. 17 (2), 60–68. doi:10.1016/j.tics.2012.12.001.

Sloman, S.A., 1996. The empirical case for two systems of reasoning. Psychol. Bull. 119 (1), 3–22.

Spielberger, C.D., Gorsuch, R.L., Lushene, R.E., 1970. STAI Manual for the State-Trait Anxiety Inventory. Consulting Psychologists Press.

Stachenfeld, K.L., Botvinick, M.M., Gershman, S.J., 2017. The Hippocampus as a Predictive Map. BioRxiv doi:10.1101/097170.

Strange, B.A., Witter, M.P., Lein, E.S., Moser, E.I., 2014. Functional organization of the hippocampal longitudinal axis. Nat. Rev. Neurosci. 15, 655–669.

Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning: An introduction. MIT Press.

Vikbladh, O.M., Meager, M.R., King, J., Blackmon, K., Devinsky, O., Shohamy, D., Burgess, N., Daw, N.D., 2019. Hippocampal contributions to model-based planning and spatial memory. Neuron 102 (3), 683–693. doi:10.1016/j.neuron.2019.02.014.

Vogel, S., Fernández, G., Joëls, M., Schwabe, L., 2016. Cognitive adaptation under stress: a case for the mineralocorticoid receptor. Trends Cogn. Sci. 20 (3), 192–203. doi:10.1016/j.tics.2015.12.003.

Vogel, S., Schwabe, L., 2016. Learning and memory under stress: implications for the classroom. Npj Sci. Learn. (1) 1. doi:10.1038/npjscilearn.2016.11.

Watanabe, S., 2010. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. J. Mach. Learn. Res. 11, 3571–3594.

Weger, M., Sandi, C., 2018. High anxiety trait: a vulnerable phenotype for stress-induced depression. Neurosci. Biobehav. Rev. 87, 27–37. doi:10.1016/j.neubiorev.2018.01.012.

Wikenheiser, A.M., Schoenbaum, G., 2016. Over the river, through the woods: cognitive maps in the hippocampus and orbitofrontal cortex. Nat. Rev. Neurosci. 17 (8), 513–523. doi:10.1038/nrn.2016.56.

Wirz, L., Bogdanov, M., Schwabe, L., 2018. Habits under stress: mechanistic insights across different types of learning. Curr. Opin. Behav. Sci 20, 9–16. doi:10.1016/j.cobeha.2017.08.009.

Zarr, N., Brown, J.W., 2016. Hierarchical error representation in medial prefrontal cortex. Neuroimage 124, 238–247. doi:10.1016/j.neuroimage.2015.08.063.

Zeidman, P., Maguire, E.A., 2016. Anterior hippocampus: the anatomy of perception, imagination and episodic memory. Nat. Rev. Neurosci. 17 (3), 173–182. doi:10.1038/nrn.2015.24.

SUPPLEMENTAL MATERIAL

Stress reduces both model-based and model-free neural computations during flexible learning

Anna Cremer[1], Felix Kalbe[1], Jan Gläscher[2,3] and Lars Schwabe[1,3]

[1] Department of Cognitive Psychology, Institute of Psychology, University of Hamburg, 20146 Hamburg, Germany

[2] Institute for Systems Neuroscience, University Medical Center Hamburg-Eppendorf, 20246 Hamburg, Germany

[3] these authors contributed equally

Corresponding author: Prof. Dr. Lars Schwabe

University of Hamburg

Department of Cognitive Psychology

20146 Hamburg, Germany

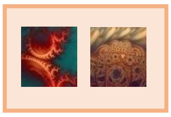
e-mail: Lars.Schwabe@uni-hamburg.de

phone: +49-40-42838-5950

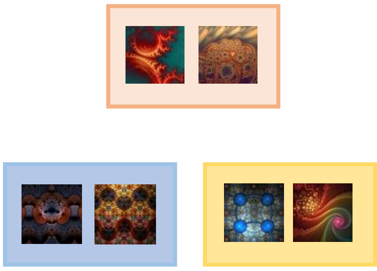
fax: +49-40-42838-4729

S1. Task instructions.

Hello! Thank you very much for participating in our experiment. In the following task, you will have the option to decide between two pictures, twice in a row. The second decision can lead to a reward. Your task is to find out which picture leads to the reward. Press space to see an example.



In the first step, you can decide between two pictures. The chosen picture leads to one of two possible environments (blue or yellow).



Now it’s time for you second decision. Again, you have to choose between two pictures. Every picture within this second decision leads to a reward, with a certain probability. Both in the blue and in the yellow environment you will find a rather advantageous and a rather disadvantageous option.

If your decision lead to the reward, a coin appears on the screen. If you did not win, the phrase “no reward” appears. The goal of this task is to find out which picture leads to the reward with the highest probability.

However, the correct answer will change several times during the experiment. As soon as you notice such a change, you should adapt your answer accordingly to continue earning rewards.

One important information is that there is no answer that always leads to a reward. This means that you can also chose the same answer that lead to a reward before without getting a reward this time. This still does not necessarily mean that a reversal took place. One picture will be rewarded almost every time, one leads to a reward often, one is sometimes rewarded and one rarely.

Do you have any questions?

Press space as soon as you are ready to start.

In order to introduce you to the task, you will now start with three phases of training. In the first phase, you will learn to react to the pictures and to find out which picture leads to the reward. Are you ready? Press the space bar to start.

[10 trials in which one option deterministically leads to a reward while the other three second stage options do not]

In this second phase you will learn to react to a reversal of the correct answer. At first, you should find out which answer leads to the reward, just as in phase 1. After some time, the correct answer will change. Please adapt your answer according to this change.

[10 trials in which one option deterministically leads to a reward while the other three second stage options do not. This option is being reversed]

The following phase is equivalent to the procedure of the actual experiment. As before, the first step is to find which picture leads to the reward. As you know, there is one more difficulty: pictures lead to a reward eventually, but they differ in the probability to actually be rewarded.

There is no picture which always results in a reward. Therefore, it is also possible, that a picture that has been successful so far, does not lead to a reward, which still does not mean that a reversal took place. Nevertheless, in this phase there will still be a reversal at some point. Try to detect the change as fast as possible and continue earning rewards!

[10 trials with the actual reward probabilities and a reversal]

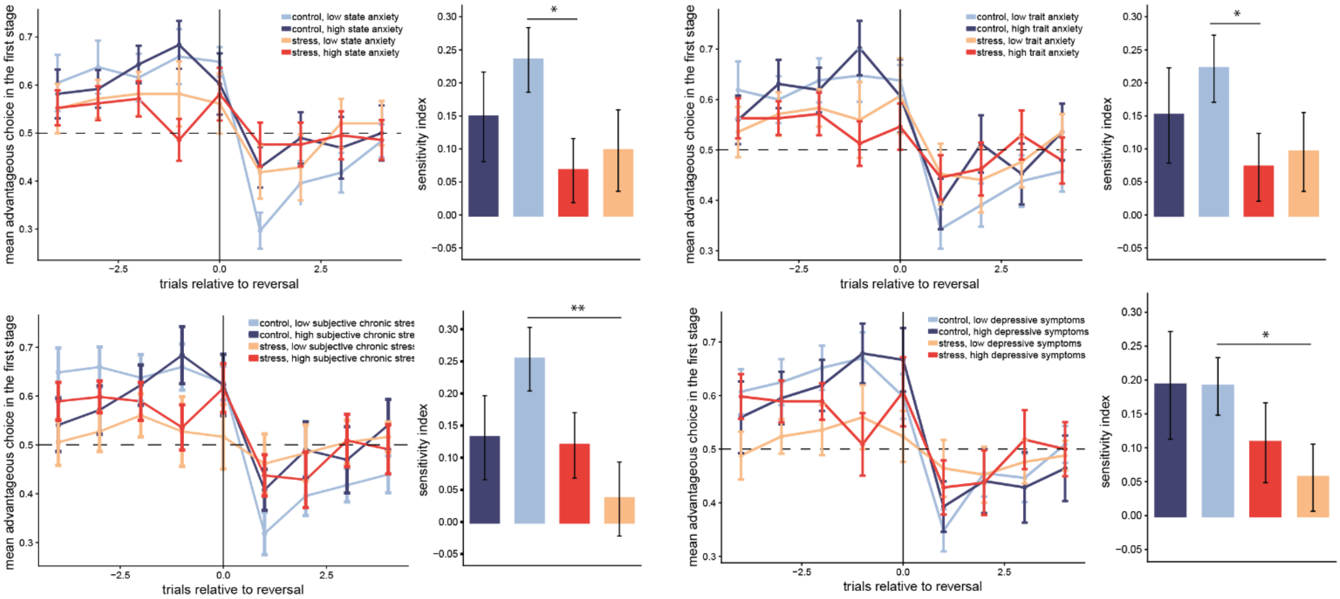


Figure S1. Effect of stress on reversal sensitivity, shown separately for high vs. low levels of anxiety, chronic stress and depressive symptoms, respectively. Acute stress appeared to influence participants' behavioral response to the reversal in particular in individuals with high trait or state anxiety. Moreover, stress and control groups appeared to differ in particular, when participants reported low chronic stress and low levels of depressive mood. Error bars represent standard errors of the mean, **p < 0.01 and *p < 0.05.

Table S1. Personality traits, behavioral tendencies and working memory capacity

	stress group	control group	difference between stress and control groups			DV = Mean advantageous choices treatment × characteristic × time		
			t	p	d	F	p	η^2_{ges}
BDI	5.79 ± 5.16	3.86 ± 3.63	1.62	0.11	0.43	0.18	0.67	0.002
STAI-S	36.10 ± 7.13	35.52 ± 5.96	0.33	0.74	0.089	0.24	0.63	0.003
STAI-T	36.79 ± 8.78	34.30 ± 7.14	1.16	0.25	0.31	0.17	0.68	0.002
TICS	13.69 ± 9.30	11.67 ± 7.45	0.89	0.38	0.24	3.33	0.07	0.04
n-back	0.69 ± 0.17	0.76 ± 0.18	-1.38	0.17	-0.38	0.89	0.35	0.01

Table S1. Control variables. Questionnaire and n-back data are mean ± standard deviation. BDI, Beck depression Inventory; STAI, State Trait Anxiety Inventory; TICS, Trier Inventory of Chronic Stress. The n-back task assessed participants' working memory capacity. The ANOVA represents the proportion of advantageous choices as a function of treatment (stress vs. control), characteristic (high value vs. low value group) and time (pre vs. post reversal).

Table S2. Correlation between control variables and model parameters.

	correlation with parameter						
	α_1	α_2	β_1	β_2	λ	p	w
BDI	-0.20	-0.20	0.04	0.08	-0.22	-0.04	-0.20
	0.14	0.14	0.77	0.56	0.10	0.79	0.89
STAI-S	-0.005	0.07	-0.11	0.004	-0.02	-0.08	-0.29
	0.97	0.59	0.40	0.98	0.89	0.56	0.83
STAI-T	-0.14	-0.16	-0.01	0.002	-0.11	-0.15	-0.18
	0.29	0.24	0.92	0.99	0.42	0.26	0.19
TICS	0.11	-0.01	-0.17	-0.18	0.02	-0.12	-0.009
	0.41	0.92	0.20	0.18	0.90	0.39	0.94
n-back	-0.17	-0.26	0.30	0.26	-0.03	-0.10	-0.08
	0.23	0.06	0.03*	0.06	0.82	0.49	0.56

Table S2. Correlation between control variables and model parameters. Data represent correlation between the questionnaire and working memory scores and the fitted model-parameters, expressed by the correlation coefficient r in the upper row and the significance level p below. BDI, Beck depression Inventory; STAI, State Trait Anxiety Inventory; TICS, Trier Inventory of Chronic Stress. The n-back task assessed the working memory capacity, * $p < 0.05$.

Appendix B: Study 2

Cremer, A., Kalbe, F., Müller, J. C., Wiedemann, K., & Schwabe, L. (2023). Disentangling the roles of dopamine and noradrenaline in the exploration-exploitation tradeoff during human decision-making. Neuropsychopharmacology, 48(7), 1078-1086. https://doi.org/10.1038/s41386-022-01517-9

www.nature.com/npp

American College *of* Neuropsychopharmacology

ARTICLE OPEN

Disentangling the roles of dopamine and noradrenaline in the exploration-exploitation tradeoff during human decision-making

Anna Cremer[1], Felix Kalbe[1], Jana Christina Müller[2], Klaus Wiedemann[2] and Lars Schwabe[1]✉

Balancing the exploration of new options and the exploitation of known options is a fundamental challenge in decision-making, yet the mechanisms involved in this balance are not fully understood. Here, we aimed to elucidate the distinct roles of dopamine and noradrenaline in the exploration-exploitation tradeoff during human choice. To this end, we used a double-blind, placebo-controlled design in which participants received either a placebo, 400 mg of the D2/D3 receptor antagonist amisulpride, or 40 mg of the β-adrenergic receptor antagonist propranolol before they completed a virtual patch-foraging task probing exploration and exploitation. We systematically varied the rewards associated with choice options, the rate by which rewards decreased over time, and the opportunity costs it took to switch to the next option to disentangle the contributions of dopamine and noradrenaline to specific choice aspects. Our data show that amisulpride increased the sensitivity to all of these three critical choice features, whereas propranolol was associated with a reduced tendency to use value information. Our findings provide novel insights into the specific roles of dopamine and noradrenaline in the regulation of human choice behavior, suggesting a critical involvement of dopamine in directed exploration and a role of noradrenaline in more random exploration.

Neuropsychopharmacology (2023) 48:1078–1086; https://doi.org/10.1038/s41386-022-01517-9

INTRODUCTION

During choice, we often face the difficult decision of when to leave a known option in favor of a potentially better, but unknown alternative. While the exploitation of a known option comes with a predictable immediate reward, exploring new options is associated with a potentially higher payoff but also the risk of a low(er) reward. At the same time, exploration provides information for improving future decisions [1–3]. Extensive exploitative behavior is further linked to inflexibility and may impede gathering new information about the environment, while an extensive exploration may lead to inefficient and inconsistent decision-making, thus reducing long-term payoffs [4, 5]. Consequently, a successful adaption to complex and volatile environments requires an intricate balance of exploration and exploitation. Biases in the exploration-exploitation tradeoff have been associated with psychiatric disorders, such as addiction [6], gambling disorder [7], or anxiety disorder [8]. Given the fundamental relevance of the exploration-exploitation trade-off for adaptive behavior, understanding the mechanisms through which humans and other animals balance exploration and exploitation during decision-making is crucial.

Neural data suggest that exploration and exploitation rely on distinct brain systems, with exploitation being associated with a mechanism in the ventromedial prefrontal cortex (vmPFC) [9, 10] while exploration is linked to a track from the frontopolar cortex to the lateral PFC [2, 11, 12]. Importantly, there is accumulating evidence that exploration and exploitation not only rely on distinct neural circuits but that these processes might also be characterized by a differential involvement of major neurotransmitters, namely dopamine and noradrenaline. Striatal dopamine is commonly associated with signaling reward values and predicting future rewards [13–15]. In line with these findings, genes involved in striatal dopamine signaling were linked to exploitation [16]. However, there is also evidence suggesting a key role of dopamine in explorative behavior, associated with genes implicated in prefrontal dopamine function. Participants with a variation of the cathecol-O-methyltransferase (COMT) gene – associated with higher tonic levels of dopamine – made exploratory decisions in proportion to the uncertainty about whether alternative options might lead to better outcomes than the status quo [16]. One potential mechanism that may underlie this so-called ‘directed’ exploration is a novelty bonus that is added to unknown alternatives and may promote the acquisition of new information [17]. In line with the idea that dopamine plays a role in directed exploration, novel stimuli excite dopaminergic neurons and activate brain regions receiving dopaminergic input [18, 19].

Noradrenaline has also been repeatedly associated with exploratory behavior. For instance, high levels of noradrenaline have been shown to increase the probability of strategy shifts, whereas low levels of noradrenaline facilitate perseverative behavior [20]. In sharp contrast to dopamine, however, noradrenaline appears not to induce a bias towards information seeking when facing uncertainty (i.e., directed exploration), but rather to

[1]Department of Cognitive Psychology, Universität Hamburg, Hamburg, Germany. [2]Department of Psychiatry and Psychotherapy, University Medical Center Hamburg- Eppendorf, Hamburg, Germany. ✉email: lars.schwabe@uni-hamburg.de

Received: 28 March 2022 Revised: 29 November 2022 Accepted: 30 November 2022
Published online: 15 December 2022

promote so-called ‘random exploration’ in which the induction of stochasticity leads to a value-independent exploration. Specifically, rodent studies showed that boosting noradrenaline leads to more value-free-random-like random behavior [21], whereas a pharmacological blockade of noradrenaline in monkeys resulted in increased choice consistency [22]. Noradrenaline might exert these effects by acting as a ‘reset button’ that interrupts ongoing information processing [20], thereby inhibiting the use of previously accumulated knowledge in favor of exploring new options [23].

Understanding the exact roles of dopamine and noradrenaline in the exploration-exploitation tradeoff may aid the development of new tools enabling the modulation of this tradeoff. However, to date, the distinct roles of dopamine and noradrenaline in the exploration-exploitation balance are not fully understood. Thus, the present experiment aimed to elucidate the specific roles of dopamine and noradrenaline in the exploration-exploitation tradeoff in human choice. We disentangled the involvement of dopamine and noradrenaline in specific sub processes underlying exploration and exploitation in a virtual patch-foraging task, which has been used before to dissociate exploration, operationalized as patch switching, and exploitation processes [24, 25]. Specifically, we systematically manipulated the rewards associated with the choice options, the degree to which the reward decreased, and the time it took to get to the next option. The degree to which these variables affect participants’ choice behavior may indicate to which extent explorative behavior is directed or more random.

MATERIALS AND METHODS

Participants and experimental design

Sixty-nine healthy volunteers (33 women, 36 men) between 18 and 35 years of age (mean = 24.98, sd = 3.67) were pseudorandomly assigned to one of three groups, controlling for a comparable gender allocation across groups: placebo (n = 22, 10 women), amisulpride (*n* = 23, 11 women) or propranolol (*n* = 24, 12 women). This sample size was based on a previous study examining the effect of amisulpride and propranolol on cognitive processing [26]. A-priori power analysis using G*Power [27] indicated that a sample of 63 participants is required in order to detect an effect a medium to large effect – as reported in [26]– with a power of 0.95. Because we expected a drop-out rate of up to 10 percent, we aimed at a sample size of 69 participants. Individuals with a current medical condition, current medication intake, lifetime history of any neurological or psychiatric disorder, drug or tobacco use, or intake of hormonal contraceptives in women (in order to avoid interactions with the administered drugs) were excluded from participation. Participants were further asked to refrain from caffeinated beverages and not to do any exercise on the day of the experiment. In addition, they should not eat or drink anything except water 2 h before the appointment. All testing took place in the afternoon and early evening, with the time of testing being counterbalanced across groups. All participants provided written informed consent before the beginning of the appointment and received a moderate monetary compensation. The study protocol was approved by the ethics committee of the Medical Chamber of Hamburg (PV7044).

Pharmacological treatment

To determine the role of noradrenaline and dopamine in the exploration-exploitation tradeoff during human choice, we used a placebo-controlled, double-blind, between-subject design in which participants received orally either a placebo, 40 mg of the β-adrenoceptor antagonist propranolol, or 400 mg of the dopaminergic D2/D3 receptor antagonist amisulpride. The dosages of the drugs were based on previous studies on the role of noradrenaline and dopamine, respectively, in cognitive processes [28–31]. Because of the distinct pharmacokinetics of propranolol and amisulpride, and in line with previous studies [23, 26, 32], we administered these drugs at two separate time points. Amisulpride was administered 120 min, and propranolol 90 min before task onset. All participants received a pill at both time points, with the amisulpride group obtaining amisulpride at the first time point, followed by a placebo at the second time point and the propranolol group receiving first a placebo and subsequently propranolol. The placebo group received a placebo at both time points. Pills were indistinguishable both for the participants and the experimenter (double-blind). Participants’ intake of the pills was monitored by an experimenter.

To verify the action of the drugs, we measured blood pressure and heart rate at several time points before and after drug administration (at baseline and 90, 120, 150 and 180 min after intake of the first pill, see Fig. 2) using a digital device (OMRON model M500 (HEM-7321-D); Healthcare Europe BV, Hoofddorp, The Netherlands) with a cuff applied around the right upper arm, when participants were sitting. We took two measures (~45 s), with a 30 s interval in between. We took the raw data provided by the device and used the mean of the two measurements per time point for the manipulation check. Moreover, we measured pupil diameter and blink rate using a RED-m eyetracker (SensoMotoric Instruments GmbH) at baseline (T_1) and 90 min after the first pill was administered (T_2). At both time points, participants were asked to fixate a black cross, presented centrally on a gray background, for 60 s. At the beginning of the measurements, each participant’s point-of-gaze was calibrated using a 5-point calibration sequence provided by the SMI software. The software automatically returned the number of blinks counted within the 60 s and the mean pupil diameter (in mm) within this period. We did not further process the data. Changes in blink rate were quantified by the number of blinks during fixation time at T_2 minus T_1, and changes in pupil size were assessed by the pupil diameter at T_2 minus $T_{1.}$

Foraging task

Participants performed a sequential patch-foraging task that had been used previously to dissociate explorative and exploitative behavior [24, 25]. Participants visited virtual orchards where they had to harvest apple trees with the goal to collect as many apples as possible within a limited amount of time. On each trial, they had to decide whether to stay at the current tree and harvest, or to move to the next tree (see Fig. 1). Patch switching was taken as an indicator of exploration. Each subsequent harvest of the same tree resulted in a slightly decreased return, so that at some point it was advantageous to move to the next tree. In addition to the expected reward, we manipulated the time required to reach the next tree (travel time) which was assumed to play a key role in the decision whether to continue harvesting the current tree or moving to the next tree. Travel time could be either 6 s (short) or 12 s (long) and was stable within an orchard. Participants performed four blocks, each for a fixed time of 7 min, resulting in a total task duration of 28 min. Blocks with short and long travel time orchards were alternating. Whether participants started with the short or the long travel time orchard was counterbalanced across participants and groups. The difference in travel time was used as a switching cost with switching being less advantageous in long travel times, because no apples could be collected during this time.

On each trial, participants submitted their choice via button press, using the down arrow for harvesting the currently displayed tree and the right arrow for moving on to the next tree. A white dot appeared under the tree indicating that a decision should be placed. If the participant decided to harvest the tree, the number of harvested apples was displayed after a

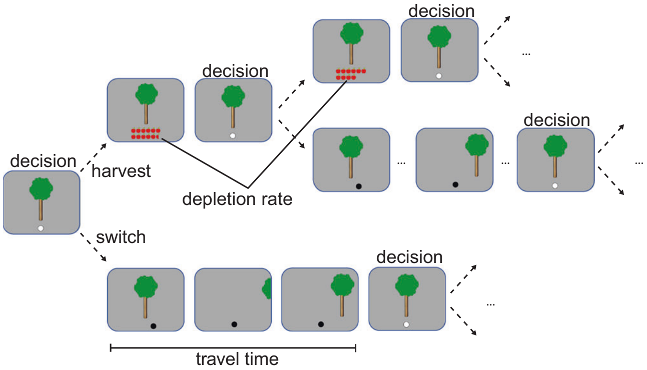


Fig. 1 Experimental task. On each trial, participants choose whether to stay at the current tree and harvest or to switch to the next tree. Each subsequent harvest of the same tree resulted in a slightly decreased outcome and switching comes with the cost of travel time. Initial richness of trees and depletion rate differed between trees, but were equally distributed in environments with long and short travel times, respectively.

harvest time of three seconds, followed by the white dot asking for the next decision. If the participant chose to switch to the next tree, the dot turned black and the way to the next tree was displayed, either for 6 s or for 12 s, depending on the environment.

Decisions had to be placed within 1 s, otherwise a warning appeared, followed by a short timeout before the next decision could be submitted. With each repeated harvest of the same tree, the yield of the tree decreased by a depletion rate. Each tree's richness, i.e., the number of apples obtained from the first harvest, was randomly drawn from a Gaussian distribution with a mean of 10 and SD of 1. The depletion rate for each successive harvest of a tree was randomly drawn from a Beta distribution with parameters 14.9 and 2.0. Participants were informed that trees would vary in terms of their richness and depletion rate (i.e., some trees would be richer or poorer than others and some trees would deplete slower or faster than others), but that the trees varied in the same way across all orchards. Participants were instructed that the only factor that might change across orchards would be the time it took to travel between trees. After each block, participants could take a short break, and determine the start of the next block themselves by button press. The different blocks were distinguished by different background colors which were counterbalanced across blocks and environment types. The total number of apples harvested throughout the task was turned into payment at the end of the experiment.

Statistical analyses

To test whether the drug manipulation was successful, blood pressure and heart rate measurements as well as eye-tracking data were analyzed using mixed-effect ANOVAs with the between-subjects factor group and the within-subject factor time. Post-hoc t-tests were used to follow-up on group differences in these measures. A mixed-effects logistic regression analysis was used to explain choice behavior. Choice was coded as stay vs. switch, indicated by 0 and 1, respectively. It was explained as a function of previous return (number of apples obtained from the previous harvest), travel time (short = 0 vs. long = 1), depletion rate, number of previous stays at current tree, and group (placebo vs. amisulpride vs. propranolol) with the placebo group as reference. We used the Akaike Information Criterion (AIC) [33] for model selection, and likelihood-ratio tests to compare our full model to gradually reduced versions. We started with a model that solely included the factor previous return and then incrementally added the factors travel time, depletion rate, number of previous stays, and group. The final model contained these five predictors, and their interaction with the experimental group (except for the factor group itself). All models consisted of the factor(s) as fixed effect(s), the overall intercept, and a random intercept per subject.

In a next step, we tested whether the factors' estimates changed over time and whether this was different in the experimental groups. Therefore, we fitted our model separately for the first half of the task (blocks 1 and 2) and the second half (blocks 3 and 4). Note that a blockwise comparison cannot be applied here, since the blocks had either an environment with short or long travel time and these blocks were alternating. Whether the first block contained a short or long travel time orchard was counterbalanced so that an analysis based on continuous blocks would compare choices at short travel times to behavior at long travel times.

To further quantify task performance, we tested whether the total sum of rewards obtained throughout the task and the proportion of switch choices differed between the experimental groups in ANOVAs with the between factor group. In a next step, we tested whether the task performance measures differed in environments with short versus long travel times in mixed-effect ANOVAs with the between-subjects factor group and the within-subject factor travel time. All analyses were performed in R [34]. Greenhouse-Geisser correction was applied when sphericity was violated. Logistic regressions were conducted as mixed-effects models and were performed using the *lme4* package [35].

Marginal value theorem

In an exploratory analysis, we applied the marginal value theorem (MVT) which describes the optimal behavior in patch-foraging decisions. Although the purpose of our study was not to assess whether participants used an optimal strategy, but to examine group differences in the use of information given by the task, the MVT may provide additional insights into participants behavior. Originally stated in animal literature, it assumes that an individual should leave the current option when the return falls below the average return in the environment [36]. Therefore, the optimal strategy is to switch when the expected number of apples to be obtained at the next harvest falls below the average return in the current environment:

$$\mathbb{E}[r_{i+1}] < \rho h \tag{1}$$

The immediate expected reward $\mathbb{E}[r]$ in the upcoming trial $i + 1$ results from reward in the current trial r, discounted by the depletion rate κ. The average return in the environment is reflected by the overall richness of the environment per timestep, i.e., the average reward in the current environment ρ multiplied by the harvest time h. Consequently, the MVT states that the maximum reward is yielded when participants switch at:

$$\kappa r < \rho h \tag{2}$$

Therefore, ρh is the threshold at which the participant should leave the current tree in favor for a new option. We simulated the optimal theshold for our task by modeling the task structure and entering all possible leaving thresholds, then probabilistically returning the expected reward over time for each threshold. We used the *optimize* function from the *stats* package in R [34] to find the exit threshold that leads to the maximum number of rewards, separately for environments with short and long travel times. For the short travel time environment this threshold is 6.7, for the long travel time environment it is 5.67. We then determined each participant's individual leaving threshold by averaging the number of apples harvested in the last two trials before leaving to the next tree. We excluded cases in which a tree was only harvested once [25]. We used t-tests to check whether the exit thresholds in the experimental groups significantly deviated from the optimal thresholds. Further, we tested whether the exit thresholds for each environment differed between groups in an ANOVA with the between factor group.

Computational modeling

We fitted an MVT model to our data using an error driven learning algorithm for the difference $\kappa r–\rho h$ [24]. The model contains a learning rate α, an inverse temperature parameter β, and an intercept c. The average reward rate in the current environment ρ was updated trial-by-trial according to the difference between the actual and the expected reward δ, and weighted by a learning rate α. Note that the prediction error δ refers to the reward per timestep, therefore includes the time τ passing in the corresponding trial (harvest time h for stay choices, travel time d for switch choices):

$$\delta = \frac{r_i}{\tau_i} - \rho_i \tag{3}$$

ρ is updated by:

$$\rho_{i+1} = \rho_i + \left[1 - (1 - \alpha)^{\tau_i}\right] \cdot \delta_i \tag{4}$$

resulting in:

$$\rho_i = (1 - \alpha)^{\tau_i} \frac{r_i}{\tau_i} + \left[1 - (1 - \alpha)^{\tau_i}\right] \rho_{i-1} \tag{5}$$

The probability P for the action a_i was derived by the choice rule:

$$P(a_i = harvest) = 1/\{1 + \exp[-c - \beta(\kappa_k r_i - \rho_i h)]\} \tag{6}$$

The learning rate α indicates the degree to which a prediction error leads to an adjustment of action values. It is constrained from 0 to 1 with higher values indicating a higher influence of δ. The inverse temperature parameter β, ranging from 0 to ∞, reflects the extent to which the action values influence choice. Higher β values stand for more value dependent choice behavior, i.e., participants choose the option with the highest expected value, while low β parameters indicate value indepentent choices, i.e., random behavior. The intercept c can reach values from 0 to ∞ and captures any constant choice biases with higher values indicating a bias towards staying and lower values representing a bias towards switching. Please see [24] for model proof and further details. Each participant's best fitting parameters were estimated by maximum likelihood estimation using the *optim* function in the *stats* package [34].

RESULTS

Manipulation check

To confirm the action of the drugs, we assessed changes in blood pressure, heart rate, blink rate and pupil diameter. Heart rate

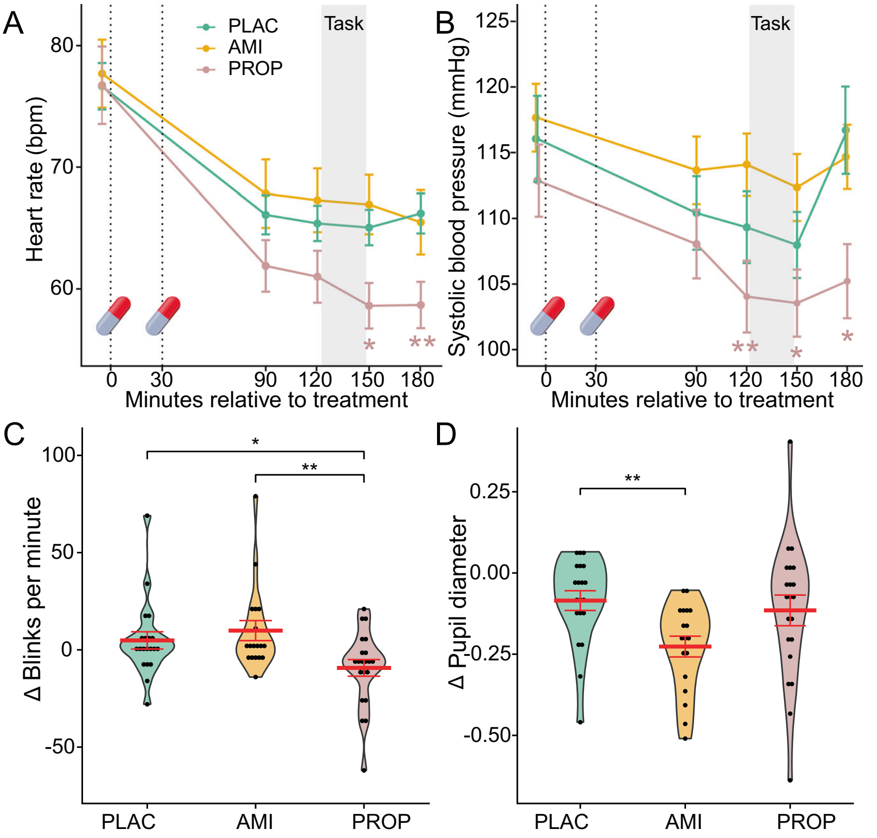


Fig. 2 Manipulation check. The action of the drugs was confirmed by physiological measures. Heart rate and blood pressure decreased in all participants across the experiment, however, significantly more pronounced in the propranolol group, compared to the placebo group and the amisulpride group (**A**, **B**). The blink rate decreased after propranolol intake, but not after both placebo and amisulpride (**C**). A decrease of the pupil diameter was present in all groups with the strongest effect in the amisulpride group (**D**). Eye-tracking data were assessed at baseline (T_1) and 90 min after the first pill intake (T_2), change values reflect T_2-T_1. Plots show binned individual data, the width corresponds to the bin's frequency (**C**, **D**); error bars represent the standard errors of the mean, *$p < .05$, **$p < .01$ for the post hoc t-tests between the groups.

decreased in all participants across the experiment, however, significantly more pronounced in the propranolol group than in the other two groups (treatment×time: $F(5.05, 164.09) = 3.12$, $p = 0.010$ (Greenhouse-Geisser corrected), $\eta^2_{ges} = 0.01$, Fig. 2A). Shortly before the foraging task, heart rate tended to be lower in the propranolol group, compared to both the placebo group ($t(43) = -1.68$, $p = 0.099$, $d = -0.50$) and the amisulpride group ($t(44)$ -1.86, $p = 0.069$, $d = -0.55$). Immediately after the task, heart rate was significantly lower in the propranolol group than in the placebo ($t(43) = -2.70$, $p = 0.010$, $d = -0.50$) and amisulpride groups ($t(44) = -2.70$, $p = 0.010$, $d = -0.55$).

Similarly, systolic blood pressure decreased significantly more strongly in the propranolol group than in the placebo and amisulpride groups (time×group: $F(6.43, 208.89) = 2.91$, $p = 0.008$ (Greenhouse-Geisser corrected), $\eta^2_{ges} = 0.1$; diastolic blood pressure: time×group: $F(7.04, 228.76) = 1.21$, $p = 0.30$ (Greenhouse-Geisser corrected), $\eta^2_{ges} = 0.008$). Systolic blood pressure was significantly lower in the propranolol group than in the amisulpride group immediately before and after the foraging task (120 min after baseline: $t(44) = -2.78$, $p = 0.008$, $d = -0.82$; 150 min after baseline: $t(44) = -2.44$, $p = 0.019$, $d = -0.72$, and 180 min after baseline: $t(44) = -2.53$, $p = 0.015$, $d = -0.75$; Fig. 2B). Compared to the placebo group, systolic blood pressure was also lower in the propranolol group, this difference, however, was significant only 180 min after pill intake ($t(43) = -2.64$, $p = 0.011$, $d = -0.79$).

Blink rate differed between groups ($F(2, 56) = 4.73$, $p = 0.013$, $\eta^2_{ges} = 0.14$) with a significant decrease from baseline to pre-task in the propranolol group, compared to placebo ($t(39 = -2.29$, $p = 0.027$, $d = -0.72$) and amisulpride ($t(37) = -2.89$, $p = 0.006$, $d = -0.93$; Fig. 2). Likewise, the pupil dilation differed between groups, but in contrast to the cardiovascular measures and the blink rate, it changed particularly after amisulpride intake ($F(2, 56) = 3.64$, $p = 0.033$, $\eta^2_{ges} = 0.12$). As shown in Fig. 2D, pupil dilation showed a significantly stronger decrease in response to amisulpride intake, compared to placebo ($t(36) = -3.20$, $p = 0.003$, $d = -1.04$), and a tendency to a more pronounced decline in contrast to the propranolol group ($t(37) = -1.89$, $p = 0.067$, $d = -0.61$), in line with previous evidence showing an impact of amisulpride, but not propranolol [37], on pupil dilation [38]. To test whether the peripheral drug effects confounded our results, we tested whether changes in blood pressure and eye-tracking data correlated with the modeling parameters. Changes were assessed as maximum of blood pressure (systolic/diastolic) and pulse minus baseline, respectively. Changes in blink rate and pupil diameter were quantified by measures at time point 2 minus values at time point 1. None of the tests indicated an association between drug-induced changes in physiological parameters and the proportion of switch choices (all $r < |0.13|$, all $p > 0.30$), indicating that peripheral changes alone were not significantly associated with participant's choice behavior.

Distinct roles of dopamine and noradrenaline in human exploration-exploitation

In order to analyze the individual tendency to explore or exploit, we performed a mixed-effects logistic regression. This allowed us to (i) identify factors that influence choice behavior and (ii) examine whether these influences differ between groups. Choice was explained as a function of previous return, traveltime, depletion

Table 1. Model comparison by the Akaike Information Criterion (AIC).

Model	Model description	N params	AIC	χ^2	df	*p*
Model 1	Previous return	3	15550			
Model 2	Model 1 + travel time	4	15413	139.49	1	<0.001
Model 3	Model 2 + depletion rate	5	15360	55.273	1	<0.001
Model 4	Model 3 + number of previous stays	6	14928	433.96	1	<0.001
Model 5	Model 4 + group	8	14930	2.026	2	0.544
Model 6	Model 5 + interactions	16	14859	86.743	8	<0.001

The full model fitted participants' choices best in a model comparison that considers differences in model complexity. Model performance is indicated by the Akaike Information Criterion (AIC). Lower values represent a better fit. The full model contains the factors previous return, travel time, depletion rate, number of previous stay decisions for the current tree, group, and the interaction of the first four factors with the experimental group.

rate, number of previous stays, group, and the interaction of the four main factors with group. We selected this model by incrementally adding a factor and tested whether it improved the model fit, compared to the reduced version.

Separate model comparisons using likelihood ratio tests confirmed that the full model including all four main factors and their interaction with the experimental group was most appropriate. This was further reflected by the lowest (i.e., best) AIC value (Table 1).

The mixed-effect logistic regression indicated that previous reward and travel time had significant effects on choice behavior. Participants in all three groups switched less when previous returns were high (main effect of previous reward, $\beta = -0.749$, $z = -21.259$, $p = <0.001$). Importantly, however, this effect was differently pronounced in the groups. Compared to placebo, the amisulpride group switched significantly less often when previous rewards were high (previous return×amisulpride: $\beta = -0.192$, $z = -3.625$, $p < 0.001$). In sharp contrast to the amisulpride group, the propranolol group switched more often after high rewards, compared to placebo (previous return×propranolol: $\beta = 0.092$, $z = 1.956$, $p = 0.050$, Fig. 3A).

Furthermore, as expected, a long travel time was associated with less switching (main effect of travel time, $\beta = -0.691$, $z = -8.214$, $p = <0.001$). This effect, however, was more pronounced in the amisulpride group, compared to placebo (travel time×amisulpride: $\beta = -0.623$, $z = -4.948$, $p < 0.001$), indicating that the amisulpride group was particularly reluctant to switching in the face of a long travel time. The propranolol group, in turn, did not differ from the placebo group ($\beta = -0.076$, $z = -0.644$, $p = 0.520$). The depletion rate alone did not have an impact on choices, neither in the placebo group (main effect of depletion rate: $\beta = -0.287$, $z = -0.376$, $p = 0.701$), nor in the propranolol group (depletion rate×propranolol: $\beta = -0.574$, $z = -0.531$, $p = 0.595$). Interestingly, in interaction with amisulpride a higher depletion rate was associated with a higher probability to switch (depletion rate×amisulpride: $\beta = 2.685$, $z = 2.298$, $p = 0.022$, Fig. 3C).

We further tested whether the choice behavior developed throughout the task by fitting the model separately for the first half of the task (blocks 1 and 2) and for the second half (blocks 3 and 4). In general, both the results of the first and second half are in line with the overall analysis. Participants switched less when the previous return was high (first half: $\beta = -0.80$, $z = -15.49$, $p < 0.001$; second half: $\beta = -0.87$, $z = -15.81$, $p < 0.001$), and when the travel time was long (first half: $\beta = -0.56$, $z = -4.69$, $p < 0.001$; second half: $\beta = -0.91$, $z = -7.24$, $p < 0.001$). The influence of the depletion rate, however, emerged throughout the task – in the first half it did not influence choice behavior ($\beta = 0.31$, $z = 0.29$, $p = 0.77$), while in the second half participants switched even more when the depletion rate was low ($\beta = -3.16$, $z = -2.68$, $p = 0.007$). Interestingly, this analysis points towards overall behavioral biases both in the amisulpride and in the propranolol group. In the first half, the amisulpride group showed a significantly enhanced switching behavior, compared to the placebo group ($\beta = 2.36$, $z = 2.58$, $p = 0.010$), while the propranolol group did not differ from placebo ($\beta = -0.32$, $z = -0.38$, $p = 0.70$). In the second half, however, the propranolol group switched less than the placebo group ($\beta = -1.6$, $z = -1.88$, $p = 0.061$), while the amisulpride group did not differ from placebo ($\beta = 0.97$, $z = 1.06$, $p = 0.29$). Other than that, the results confirm the findings from the overall analysis: the amisulpride group switched less, when the previous return was high (first half: $\beta = -.0.28$, $z = -3.50$, $p = 0.0005$; second half: $\beta = -0.22$, $z = -2.67$, $p = 0.008$) and when the travel time was long (first half: $\beta = -1.17$, $z = -6.29$, $p < 0.0001$; second half: $\beta = -0.34$, $z = -1.85$, $p = 0.06$). Likewise, participants in the amisulpride group switched more when the depletion rate was high (first half: $\beta = 3.25$, $z = 1.93$, $p = 0.05$; second half: $\beta = 3.64$, $z = 2.06$, $p = 0.04$, Supplementary Fig. S2 in the Supplementary Material). Again, neither of the choice factors significantly influenced decision making in the propranolol group.

Task performance

Groups did not differ in the number of total rewards obtained throughout the task ($F(2,66) = 1.68$, $p = 0.19$, $\eta^2_{ges} = 0.048$). However, participants in the amisulpride group tended to collect more rewards, compared to placebo ($t(43) = 1.92$, $p = 0.061$, $d = 0.57$) and propranolol ($t(45) = 1.65$, $p = 0.11$, $d = 0.48$). The number of rewards differed between environments with short and long travel time (main effect of travel time: $F(1, 66) = 229.85$, $p < 0.0001$, $\eta^2_{ges} = 0.36$, Fig. 3), but there was no significant interaction between environment and experimental group ($F(2,66) = 1.176$, $p = 0.31$, $\eta^2_{ges} = 0.006$). In environments with short travel times, the amisulpride group tended to yield higher rewards, compared to the propranolol group ($t(45) = 1.80$, $p = 0.078$, $d = 0.53$). In long travel time environments participants tended to earn more rewards after amisulpride intake than after placebo ($t(43) = 1.97$, $p = 0.055$, $d = 0.59$; all other $p > 0.15$, Fig. 3D). Overall, the groups did not differ in the percentage of switch decisions ($F(2,66) = 0.48$, $p = 0.62$, $\eta^2_{ges} = 0.14$). The percentage differed between environments with short and long travel times (main effect of travel time: $F(1, 66) = 36.89$, $p < 0.0001$, $\eta^2_{ges} = 0.056$), but this was not differentially pronounced in the experimental groups (group×travel time: $F(2, 66) = 1.03$, $p = 0.36$, $\eta^2_{ges} = 0.003$; all post hoc t-tests $p > 0.16$).

Marginal Value theorem

Exit thresholds differed between environments (main effect of travel time: $F(1,66) = 47.70$, $p < 0.0001$, $\eta^2_{ges} = 0.072$) but did not differ between groups ($F(2, 66) = 0.37$, $p = 0.69$, $\eta^2_{ges} = 0.01$). There was no group×travel time interaction ($F(2,66) = 1.27$, $p = 0.29$, $\eta^2_{ges} = 0.004$). Neither group differed from the optimal exit threshold, as supposed by the MVT (6.7 for short travel times environments, all $p > 0.76$; 5.67 for long travel time environments, all $p > 0.85$, Fig. 3E).

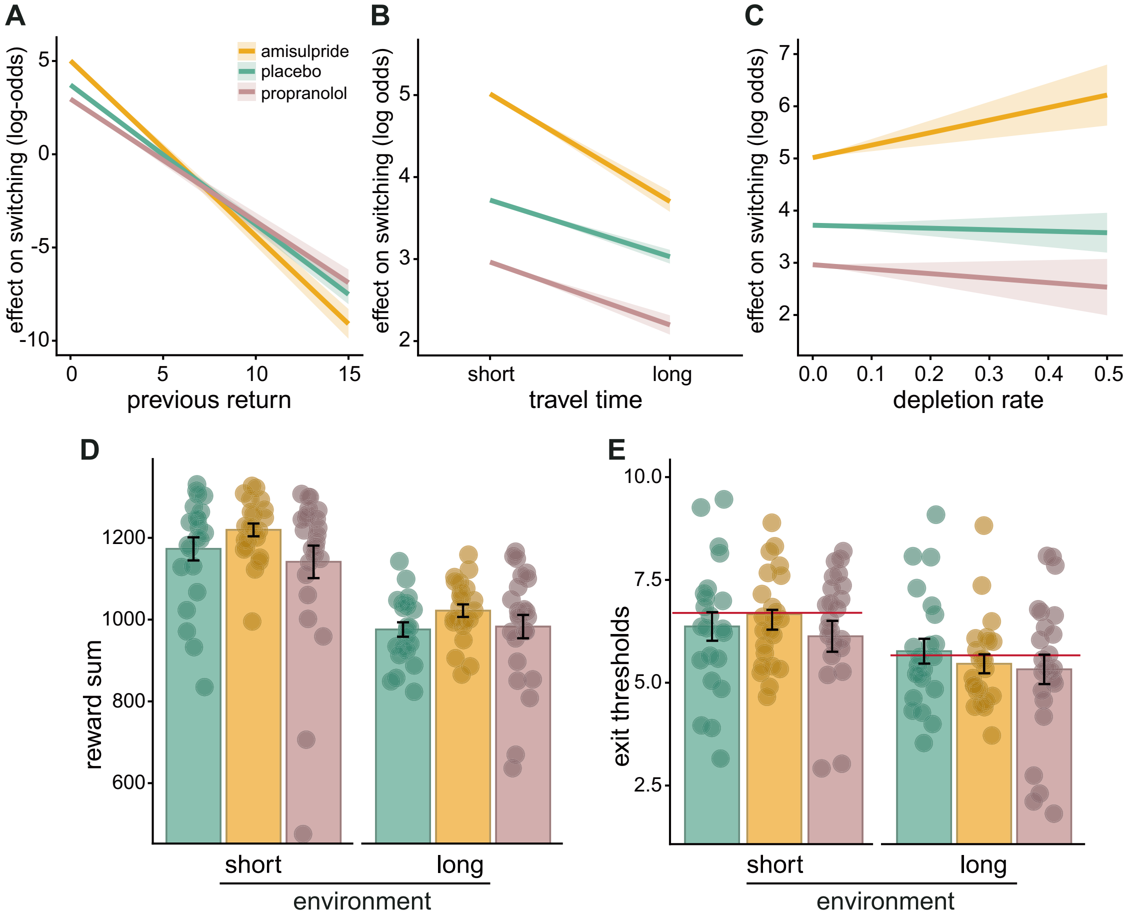


Fig. 3 Modulation of the extent to which choice features drive behavior by amisulpride and propranolol. A Participants were more likely to stay at the current option when the previous reward was high and the impact of the previous reward was particularly high after amisulpride intake. **B** The impact of travel time on choice behavior was mostly pronounced in the amisulpride group, with a lower tendency to switch in particular when the travel time was long. **C** While the depletion rate did not influence choices in placebo and propranolol groups, participants in the amisulpride group switched in particular in the face of a high depletion rate (**D**). The amisulpride group tended to collect a higher number of rewards in environments with short travel times, compared to the propranolol group (t(45) = 1.80, $p = 0.078$, $d = 0.53$). In blocks with long travel times, participants tended to collect more rewards after amisulpride intake, compared to placebo (t(43) = 1.97, $p = 0.055$, $d = 0.59$). **E** Exit thresholds differed between environments, but not between groups, red lines reflect the optimal exit threshold in the short and in the long travel time environment, following the marginal value theorem (MVT). Please note that the critical difference between the amisulpride and propranolol groups (**A**–**C**) remains if the two participants with the lowest reward sum (**D**) are removed from the analysis. Error bars represent standard errors of the mean.

Computational modeling

We fitted a computational model according to the MVT to estimate each participant's learning rate α, temperature parameter β, and choice bias c. Regarding the learning rate, we identified three participants as outlier, as they differed more than 3 standard deviations from the group's mean (one participant from each experimental group). Interestingly, participants in the amisulpride group had a significantly lower learning rate than participants in the propranolol group (t(43) = −2.16, $p = 0.036$, d = −0.65, Fig. 4A), and tended to have a lower α compared to the placebo group (t(41) = −1.99, $p = 0.054$, $d = -0.61$, Fig. 4). The learning rates did not differ between the placebo and propranolol groups (t(42) = 0.28, $p = 0.78$, $d = 0.083$).

The temperature parameter β did not differ between groups (F(2,66) = 1.53, $p = 0.22$, $\eta^2_{ges} = 0.044$, Fig. 4B). Neither the amisulpride nor the propranolol group differed significantly from the placebo group (amisulpide vs. placebo: t(43) = 1.6, $p = 0.12$, $d = 0.48$; propranolol vs. placebo: t(44) = 0.028, $p = 0.98$, $d = 0.008$; amisulpride vs. propranolol group: t(45) = 1.51, $p = 0.14$, $d = 0.44$). Likewise, the choice bias c did not differ between groups (F(2,66) = 0.51, $p = 0.6$, $\eta^2_{ges} = 0.020$, Fig. 4C). Neither the amisulpride group, nor the propranolol group differed from placebo (amisulpride vs. placebo: t(43) = 0.97, $p = 0.34$, $d = 0.29$; propranolol vs. placebo: t(44) = 0.34, $p = 0.73$, $d = 0.10$; amisulpride vs. propranolol: (t(45) = 0.67, $p = 0.51$, $d = 0.19$).

DISCUSSION

Adaptive decision-making requires an optimal balance between choosing known options and trying new paths when the environment changes or new information is required. Given the ubiquity of exploration-exploitation tradeoffs in everyday life and their potential relevance for psychopathology, understanding the mechanisms involved in this tradeoff is important. Here we investigated the specific roles of dopamine and noradrenaline in the exploration-exploitation tradeoff by pharmacological blockade of either system using propranolol and amisulpride and systematically examining the effects of reward values, depleting returns, and opportunity costs on choice behavior. The action of the administered drugs was confirmed by specific changes in blood pressure, heart rate, pupil dilation, and blink rate. As expected, (systolic) blood pressure and heart rate decreased most prominently

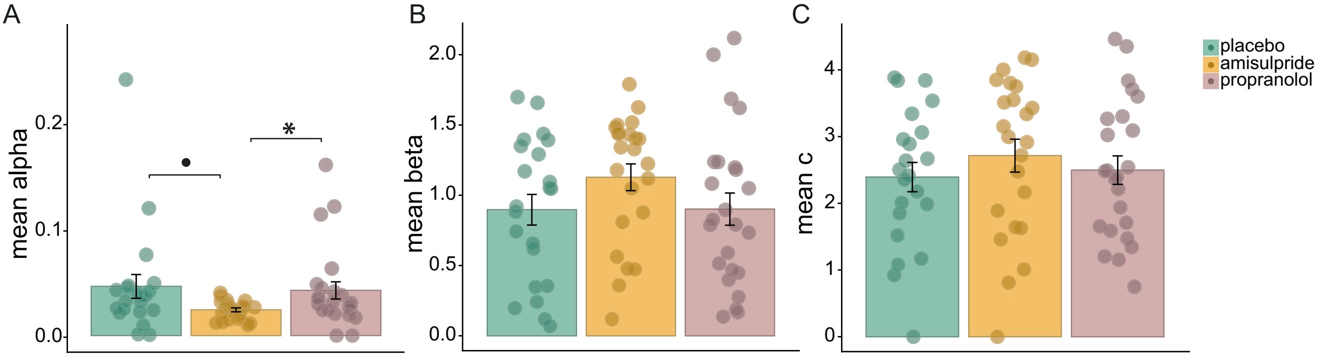


Fig. 4 Modeling parameters per subject. The learning rate α was significantly lower in the amisulpride group than in the placebo group and tended to be lower, compared to the propranolol group (**A**). Neither the temperature parameter β, nor the choice bias c differed between the experimental groups (**B**, **C**), error bars represent standard errors of the mean, * for $p < 0.05$,• for $p < 0.06$.

in the propranolol group, consistent with its action as a hypotensive agent [38], related to the blockade of β1- and β2-adrenergic receptors that represent the predominant form of adrenergic receptors expressed in the heart [39]. Propranolol was further linked to a reduced blink rate, which may be due to dryer eyes after β-adrenergic blockade [39]. Pupil diameters, in turn, known to be mediated, at least partly, by dopaminergic neurons in the ventral tegmental area (VTA; [40]) were particularly reduced in the amisulpride group, most like due to the blockade of D2/D3 receptors in the VTA [37, 41]. Most importantly, our behavioral results revealed functionally dissociable roles of dopamine and noradrenaline in the exploration-exploitation trade off, with dopamine governing the sensitivity to decision-relevant information and noradrenaline being involved in value-independent choice processes.

Previous studies suggested a role of dopamine in exploration [4, 17, 19, 42]. Our data, however, do not point towards a decrease of exploratory behavior in participants that received the D2/D3 receptor antagonist amisulpride. Instead, participants in the amisulpride group switched less, specifically when (i) the previous reward was high, (ii) the travel time was long, and (iii) the depletion rate was low. This pattern suggests an increased sensitivity to the specific choice aspects, i.e., that these had a stronger impact on choice. These results corroborate previous findings showing that D2-receptor blockade by amisulpride sharpened content-specific representations in the PFC that are used to guide reinforcement-based decisions [29, 32]. Interestingly, in the first half of the task, the amisulpride group showed significantly enhanced switching behavior, compared to the placebo group, indicating an increase in explorative choices. Taken together, these results point towards a directed exploration in the beginning of the task, which may then inform subsequent choice behavior. This is further supported by our computational modeling results. Participants in the amisulpride group had a lower learning rate compared to the other groups. Given the strong local autocorrelation of prediction errors in the present foraging task, a low learning rate may be beneficial to integrate across a longer time span. In line with our data, recent findings suggested that cabergoline, a D2 receptor agonist, reduced the sensitivity towards the difference between rich and poor environments [43]. Assuming that a D2 receptor blockade should impair dopamine-associated processes, these findings might be puzzling at first glance. However, the potential discrepancy between these findings and common beliefs about the role of dopamine in choice could be explained by a dual state model of prefrontal dopamine. This model proposes that the activation of prefrontal D1 and D2 receptors has opposing effects on GABAergic activity, resulting in bidirectional effects on the accuracy of prefrontal representations [44]. In recordings of prefrontal pyramidal neurons, a predominant D1 receptor activation (D1-dominated state) was associated with increased GABAergic inhibition, resulting in a selective access to prefrontal circuits with only very strong inputs passing through and therefore forming strong representations. A primary D2 receptor activation (D2-dominated state), on the other hand was linked to a decreased GABAergic inhibition so that multiple inputs were processed at the same time, leading to weak representations in the prefrontal cortex [44]. It is assumed that blocking prefrontal D2 receptors increases the likelihood of D1-dominated states, i.e., the processing of strong input while suppressing noise [45]. Further, amisulpride is suggested to preferably block D2/D3 receptors in the PFC, while dopamine levels in the striatum were even increased after low doses [41, 46, 47]. Our findings may thus be explained by a shift towards prefrontal D1 receptor activation in the prefrontal cortex, which may, together with an intact striatal dopamine functioning, lead to the formation of strong representations of decision-relevant stimuli and ultimately increased sensitivity for specific choice aspects to guide behavior.

In sharp contrast to the amisulpride group, none of these choice aspects had a significant effect on choice behavior in the propranolol group. Interestingly, participants in the propranolol group tended to switch even more after higher rewards, compared to the placebo group. Specifically, they still switched less after higher than lower rewards, but this was less pronounced than in the placebo group, while this effect was significantly more pronounced in the amisulpride group than in the placebo group. This pattern points to a reduced usage of decision-relevant information for choice behavior, in line with evidence suggesting a role of noradrenaline in random, but not directed exploration [21–23, 48]. However, the data on the direction of noradrenergic effects on random exploration is heterogenous. A recent study directly compared how amisulpride and propranolol affect different exploration strategies and reported that propranolol, but not amisulpride attenuated random exploration [23]. This is in line with previous findings showing that noradrenaline levels predicted increased noise in choice behavior [49]. Our data suggest an opposite effect of noradrenaline on decision noise with rather increased noise after blocking noradrenaline. The present findings dovetail with a study that reported decreased random exploration after pharmacologically elevated noradrenergic activity [48]. In the same vein, it was hypothesized that noradrenaline might work as an urgency signal that promotes commitment to an early decision. Noradrenergic blockade via propranolol was assumed to insert this signal and hence stop further information gathering [50]. This is further supported by our finding that, in the second half of the task, participants in the propranolol group showed an overall reduction of switch choices, pointing again towards a reduced use of information, but in the direction of exploitative decision-making. These heterogeneous results with respect to the direction of the influence of noradrenaline on exploration and exploitation might be related to distinct activity modes of noradrenaline. While tonic noradrenergic activity was associated with exploration, phasic noradrenaline has been thought to facilitate exploitative behavior [51]. Because there is evidence that propranolol is likely to influence both tonic and phasic signaling of noradrenaline [52], such differentiation cannot be derived from our data.

In addition to differences between tonic and phasic noradrenergic activity, a possible inhibitory mechanism of β-adrenergic receptors may explain why we found a tendency towards an increase of stochasticity. Specifically, β-adrenergic receptors enhanced inhibitory synaptic mechanisms in rats by a noradrenaline-mediated enhancement of GABA efficacy [53]. By blocking β-adrenergic receptors, we might have blocked a noradrenaline-related inhibition of noise, resulting in an increase of noisy, i.e., random behavior. This was not captured by a decreased temperature parameter in the propranolol group. However, the general range of the temperature parameter derived by the modeling approach was rather low, which can be explained by the low range in the value estimation. The temperature parameter specifies the degree to which value estimates influence behavior. Since the initial rewards were drawn from a Gaussian distribution with a mean of 10 and SD of 1, depleting by a Beta distribution with parameters 14.9 and 2.0, the estimated values came in a low range per se. Consequently, the degree to which this estimation influenced decision-making may not be suitable to interpret group differences in this case.

Overall, however, the influence of propranolol on the exploration-exploitation tradeoff was less pronounced than for amisulpride. A potential explanation for this could be that noradrenaline does not drive specific components of decision-making, but rather exerts higher-order control signals, such as an urgency signal that stops ongoing information gathering, presumably by inducing decision noise.

At this point, it should be noted that other factors such as tiredness or boredom might have affected switching behavior. Although these factors may also contribute to more random exploration and we do not think that these could explain the influence of the drugs on the dependency of switch behavior on relevant decision parameters, future studies should measure these additional variables to explicitly control for their influence. Moreover, future studies should consider including baseline measures of task performance to rule out performance differences between groups before drug administration or use a within-subject design instead of a between-subjects design.

Taken together, our findings suggest functionally dissociable roles of dopamine and noradrenaline in the exploration-exploitation tradeoff during human decision-making. Compared to placebo, participants in the amisulpride group switched less when the prospects in the current environment were still advantageous (i.e., high rewards and low depletion rates) and the costs associated with exploration were high (i.e., long travel time). After propranolol intake, participants tended to switch even more, compared to the placebo group, when the rewards in the current environment were still high. Thus, these data show that dopamine modulates the sensitivity to choice relevant aspects, while noradrenaline regulates when to disengage from the current information paths to randomly explore new options. Our results are thus generally in line with previously hypothesized roles of dopamine and noradrenaline in directed and random exploration, respectively. The present findings enhance our understanding of the differential roles of dopamine and noradrenaline in decision-making and might have relevant implications for mental disorders characterized by biases in the exploration-exploitation tradeoff.

REFERENCES

1. Cohen JD, McClure SM, Yu AJ. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. Philos Trans R Soc B Biol Sci. 2007;362:933–42.
2. Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ. Cortical substrates for exploratory decisions in humans. Nature. 2006;441:876–9.
3. Mehlhorn K, Newell BR, Todd PM, Lee MD, Morgan K, Braithwaite VA, et al. Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. Decision. 2015;2:191–215.
4. Chakroun K, Mathar D, Wiehler A, Ganzer F, Peters J. Dopaminergic modulation of the exploration/exploitation trade-off in human decision-making. ELife. 2020;9:e51260.
5. Schulz E, Gershman SJ. The algorithmic architecture of exploration in the human brain. Curr Opin Neurobiol. 2019;55:7–14.
6. Morris LS, Baek K, Kundu P, Harrison NA, Frank MJ, Voon V. Biases in the explore–exploit tradeoff in addictions: the role of avoidance of uncertainty. Neuropsychopharmacology. 2016;41:940–8.
7. Wiehler A, Chakroun K, Peters J. Attenuated directed exploration during reinforcement learning in gambling disorder. J Neurosci. 2021;41:2512–22.
8. Grupe DW, Nitschke JB. Uncertainty and anticipation in anxiety: an integrated neurobiological and psychological perspective. Nat Rev Neurosci. 2013;14:488–501.
9. Blanchard TC, Gershman SJ. Pure correlates of exploration and exploitation in the human brain. Cogn Affect Behav Neurosci. 2018;18:117–26.
10. Summerfield C, Koechlin E. A neural representation of prior information during perceptual inference. Neuron. 2008;59:336–47.
11. Donoso M, Collins AGE, Koechlin E. Foundations of human reasoning in the prefrontal cortex. Science. 2014;344:1481–6.
12. Boorman ED, Behrens TEJ, Woolrich MW, Rushworth MFS. How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. Neuron. 2009;62:733–43.
13. Bayer HM, Glimcher PW. Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron. 2005;47:129–41.
14. Lak A, Stauffer WR, Schultz W. Dopamine prediction error responses integrate subjective value from different reward dimensions. Proc Natl Acad Sci. 2014;111:2343–8.
15. Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997;275:1593–9.
16. Frank MJ, Doll BB, Oas-Terpstra J, Moreno F. Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nat Neurosci. 2009;12:1062–8.
17. Krebs RM, Schott BH, Schütze H, Düzel E. The novelty exploration bonus and its attentional modulation☆. Neuropsychologia. 2009;47:2272–81.
18. Bunzeck N, Düzel E. Absolute coding of stimulus novelty in the human substantia Nigra/VTA. Neuron. 2006;51:369–79.
19. Costa VD, Tran VL, Turchi J, Averbeck BB. Dopamine modulates novelty seeking behavior during decision making. Behav Neurosci. 2014;128:556–66.
20. Yu AJ, Dayan P. Uncertainty, neuromodulation, and attention. Neuron. 2005; 46:681–92.
21. Tervo DGR, Proskurin M, Manakov M, Kabra M, Vollmer A, Branson K, et al. Behavioral variability through stochastic choice and its gating by anterior cingulate cortex. Cell. 2014;159:21–32.
22. Jahn CI, Gilardeau S, Varazzani C, Blain B, Sallet J, Walton ME, et al. Dual contributions of noradrenaline to behavioural flexibility and motivation. Psychopharmacology. 2018;235:2687–702.
23. Dubois M, Habicht J, Michely J, Moran R, Dolan RJ, Hauser, TU. Human complex exploration strategies are enriched by noradrenaline-modulated heuristics. ELife. 2021;10:e59907.
24. Constantino SM, Daw ND. Learning the opportunity cost of time in a patch-foraging task. Cogn Affect Behav Neurosci. 2015;15:837–53.
25. Lenow JK, Constantino SM, Daw ND, Phelps EA. Chronic and acute stress promote overexploitation in serial decision making. J Neurosci. 2017;37:5681–9.
26. Hauser TU, Eldar E, Purg N, Moutoussis M, Dolan RJ. Distinct roles of dopamine and noradrenaline in incidental memory. J Neurosci. 2019;39:7715–21.
27. Faul F, Erdfelder E, Lang A-G, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods. 2007;39:175–91.
28. Burke CJ, Soutschek A, Weber S, Raja Beharelle A, Fehr E, Haker H, et al. Dopamine receptor-specific contributions to the computation of value. Neuropsychopharmacology. 2018;43:1415–24.
29. Jocham G, Klein TA, Ullsperger M. Dopamine-mediated reinforcement learning signals in the striatum and ventromedial prefrontal cortex underlie value-based choices. J Neurosci. 2011;31:1606–13.
30. Schwabe L, Römer S, Richter S, Dockendorf S, Bilak B, Schächinger H. Stress effects on declarative memory retrieval are blocked by a β-adrenoceptor antagonist in humans. Psychoneuroendocrinology. 2009;34:446–54.
31. Schwabe L, Hoffken O, Tegenthoff M, Wolf OT. Preventing the stress-induced shift from goal-directed to habit action with a -adrenergic antagonist. J Neurosci. 2011;31:17317–25.
32. Kahnt T, Weber SC, Haker H, Robbins TW, Tobler PN. Dopamine D2-Receptor blockade enhances decoding of prefrontal signals in humans. J Neurosci. 2015;35:4104–11.

33. Bozdogan H. Model selection and Akaike's Information Criterion (AIC): the general theory and its analytical extensions. Psychometrika. 1987;52:345–70.
34. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2019.
35. Pinheiro JC, Bates DM. Mixed-effects models in S and S-PLUS. New York: Springer; 2000.
36. Charnov EL. Optimal foraging, the marginal value theorem. Theor Popul Biol. 1976;9:129–36.
37. Koudas V, Nikolaou A, Hourdaki E, Giakoumaki SG, Roussos P, Bitsios P. Comparison of ketanserin, buspirone and propranolol on arousal, pupil size and autonomic function in healthy volunteers. Psychopharmacology. 2009;205:1–9.
38. Samuels ER, Hou RH, Langley RW, Szabadi E, Bradshaw CM. Comparison of pramipexole and amisulpride on alertness, autonomic and endocrine functions in healthy volunteers. Psychopharmacology. 2006;187:498–510.
39. Seal DV. The effect of ageing and disease on tear constituents. Trans Ophthalmological Societies U Kingd. 1985;104:355–62.
40. Loewenfeld IE. Mechanisms of reflex dilatation of the pupil: historical review and experimental analysis. Doc Ophthalmol. 1958;12:185–448.
41. Bressan RA, Erlandsson K, Jones HM, Mulligan R, Flanagan RJ, Ell PJ, et al. Is regionally selective D2/D3 dopamine occupancy sufficient for atypical antipsychotic effect? An in vivo quantitative [123I] epidepride SPET study of amisulpride-treated patients. Am J Psychiatry. 2003;160:1413–20.
42. Kayser AS, Mitchell JM, Weinstein D, Frank MJ. Dopamine, locus of control, and the exploration-exploitation tradeoff. Neuropsychopharmacology 2015;40:454–62.
43. Le Heron C, Kolling N, Plant O, Kienast A, Janska R, Ang Y-S, et al. Dopamine modulates dynamic decision-making during foraging. J Neurosci. 2020;40:5273–82.
44. Seamans JK, Gorelova N, Durstewitz D, Yang CR. Bidirectional dopamine modulation of GABAergic inhibition in prefrontal cortical pyramidal neurons. J Neurosci. 2001;21:3628–38.
45. Seamans JK, Yang CR. The principal features and mechanisms of dopamine modulation in the prefrontal cortex. Prog Neurobiol. 2004;74:1–58.
46. Scatton B, Claustre Y, Cudennec A, Oblin A, Perrault G, Sanger D, et al. Amisulpride: from animal pharmacology to therapeutic action. Int Clin Psychopharmacol. 1997;12:29–36.
47. Viviani R, Graf H, Wiegers M, Abler B. Effects of amisulpride on human resting cerebral perfusion. Psychopharmacology. 2013;229:95–103.
48. Warren CM, Wilson RC, van der Wee NJ, Giltay EJ, van Noorden MS, Cohen JD, et al. The effect of atomoxetine on random and directed exploration in humans. PLoS One. 2017;12:e0176034.
49. Jepma M, Nieuwenhuis S. Pupil diameter predicts changes in the exploration–exploitation trade-off: evidence for the adaptive gain theory. J Cogn Neurosci. 2011;23:1587–96.
50. Hauser TU, Moutoussis M, Purg N, Dayan P, Dolan RJ. Beta-blocker propranolol modulates decision urgency during sequential information gathering. J Neurosci. 2018;38:7170–8.
51. Aston-Jones G, Cohen JD. An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. Annu Rev Neurosci. 2005;28:403–50.
52. Lawson RP, Bisby J, Nord CL, Burgess N, Rees G. The computational, pharmacological, and physiological determinants of sensory learning under uncertainty. Curr Biol. 2021;31:163–72.e4
53. Waterhouse BD, Moises HC, Yeh HH, Woodward DJ. Norepinephrine enhancement of inhibitory synaptic mechanisms in cerebellum and cerebral cortex: mediation by beta adrenergic receptors. J Pharmacol Exp Therapeutics. 1982;221:495–506.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of Carlo Hiller with programming the task and the assistance of Roberta Souza Lima, Rosann Stocker and Fabian Schacht during data collection. This study was funded by the Landesforschungsförderung Hamburg (grant LFF FV38 to LS).

AUTHOR CONTRIBUTIONS

AC contributed to the conceptualization of the study, collected and analyzed the data, and drafted the manuscript. FK contributed to the conceptualization of the study and collected the data. JCM and KW provided medical supervision of the project. LS contributed to the conceptualization of the study, provided funds and supervision, and drafted the manuscript. All authors provided critical revisions of the manuscript.

FUNDING

Open Access funding enabled and organized by Projekt DEAL.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41386-022-01517-9.

Correspondence and requests for materials should be addressed to Lars Schwabe.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Appendix C: Study 3

Rac-Lubashevsky, R., Cremer, A., Collins, A. G., Frank, M. J., & Schwabe, L. (2023). Neural Index of Reinforcement Learning Predicts Improved Stimulus–Response Retention under High Working Memory Load. *Journal of Neuroscience*, *43*(17), 3131-3143. https://doi.org/10.1523/JNEUROSCI.1274-22.2023

Behavioral/Cognitive

Neural Index of Reinforcement Learning Predicts Improved Stimulus–Response Retention under High Working Memory Load

Rachel Rac-Lubashevsky,[1,2] Anna Cremer,[3] Anne G.E. Collins,[4,5] Michael J. Frank,[1,2*] and Lars Schwabe[3*]

[1]Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence, Rhode Island 02912, [2]Carney Institute for Brain Science, Brown University, Providence, Rhode Island 02912, [3]Department of Cognitive Psychology, Universitat Hamburg, 20146 Hamburg, Germany, [4]Department of Psychology, University of California, Berkeley, Berkeley, California 94720-1650, and [5]Helen Wills Neuroscience Institute, University of California, Berkeley, Berkeley, California 94720

Human learning and decision-making are supported by multiple systems operating in parallel. Recent studies isolating the contributions of reinforcement learning (RL) and working memory (WM) have revealed a trade-off between the two. An interactive WM/RL computational model predicts that although high WM load slows behavioral acquisition, it also induces larger prediction errors in the RL system that enhance robustness and retention of learned behaviors. Here, we tested this account by parametrically manipulating WM load during RL in conjunction with EEG in both male and female participants and administered two surprise memory tests. We further leveraged single-trial decoding of EEG signatures of RL and WM to determine whether their interaction predicted robust retention. Consistent with the model, behavioral learning was slower for associations acquired under higher load but showed parametrically improved future retention. This paradoxical result was mirrored by EEG indices of RL, which were strengthened under higher WM loads and predictive of more robust future behavioral retention of learned stimulus–response contingencies. We further tested whether stress alters the ability to shift between the two systems strategically to maximize immediate learning versus retention of information and found that induced stress had only a limited effect on this trade-off. The present results offer a deeper understanding of the cooperative interaction between WM and RL and show that relying on WM can benefit the rapid acquisition of choice behavior during learning but impairs retention.

Key words: EEG; reinforcement learning; retention; stress; working memory

Significance Statement

Successful learning is achieved by the joint contribution of the dopaminergic RL system and WM. The cooperative WM/RL model was productive in improving our understanding of the interplay between the two systems during learning, demonstrating that reliance on RL computations is modulated by WM load. However, the role of WM/RL systems in the retention of learned stimulus–response associations remained unestablished. Our results show that increased neural signatures of learning, indicative of greater RL computation, under high WM load also predicted better stimulus–response retention. This result supports a trade-off between the two systems, where degraded WM increases RL processing, which improves retention. Notably, we show that this cooperative interplay remains largely unaffected by acute stress.

Introduction

Everyday behavior, like selecting what to wear and what to eat, involves reinforcement learning (RL). Canonical RL models incrementally accumulate expected values of stimulus–action pairings over the course of multiple experiences. Although this RL system learns rather slowly and incrementally, it can be augmented by the joint support of working memory (WM), especially when learning new arbitrary contingencies (Yoo and Collins, 2021). WM enables fast learning by robustly maintaining, in an accessible form, the representations of relevant stimulus–action associations to support ongoing processing such as

Received June 29, 2022; revised Jan. 19, 2023; accepted Feb. 20, 2023.

Author contributions: M.J.F. and L.S. designed research; A.C. performed research; R.R.-L., A.C., and A.G.E.C. analyzed data; R.L.-R., A.C., M.J.F., and L.S. wrote the paper.

This work was supported by Landesforschungsfoerdung Hamburg, Germany, Grant LFF FV 38 to L.S. and National Institutes of Health Grant R01 MH084840-08A1 to M.J.F.

*M.J.F. and L.S. contributed equally to this work.

Correspondence should be addressed to Rachel Rac-Lubashevsky at rac.hunrachel@gmail.com.

https://doi.org/10.1523/JNEUROSCI.1274-22.2023

value-based learning and decision-making. However, when WM capacity is exceeded, it suffers from interference, causing relevant representations to be lost or corrupted (Oberauer et al., 2016). Indeed, although the WM system is beneficial for supporting early learning, its contribution to successful learning is constrained by limited capacity (Collins and Frank, 2012). On the other hand, the incremental RL system has a much broader capacity and is more robust as long as the reward contingencies remain stable. Previous studies have thus shown a transition from capacity- and delay-sensitive WM to RL over the course of learning (Collins and Frank, 2012, 2018).

Moreover, previous studies examining the joint contributions of WM and RL to learning have suggested that these systems are not modular but rather interactive (Collins et al., 2017a,b; Collins, 2018; Collins and Frank, 2018). fMRI and EEG studies provided support for a cooperative interaction; when stimulus–reward information is stored in WM, neural indices of reward prediction errors (RPEs) are reduced (Collins et al., 2017a; Collins and Frank, 2018). Conversely, RPEs were larger under high load, leading to accelerated neural learning curves putatively indicative of more robust RL (despite slowed behavioral learning because of degraded WM). This dissociation suggested that although a high WM load slows learning, it might also improve retention because of accumulative RPEs that reinforce the RL system. Supporting this prediction, in the surprise test phase, participants showed better retention performance for stimulus–response contingencies and their reward values when they had been learned under higher compared with lower WM demands (Collins et al., 2017b; Collins, 2018; Wimmer and Poldrack, 2022). However, two major limitations remained from this prior work.

First, the previous study showing enhanced retention of stimulus–response associations had only tested low and high WM conditions (Collins, 2018), with only subtle albeit significant differences in performance (∼5% difference between set size 3 vs 6). We thus parametrically manipulated WM demands (Collins et al., 2017b) to test the prediction that retention performance of stimulus–response associations would scale monotonically as a function of increased WM demand, despite monotonically slowed learning in these conditions. Second, although the neural and behavioral findings have been documented on their own, it has not yet been established whether cooperative neural interactions within WM/RL systems during learning are predictive of future retention. Moreover, it is unclear whether neural RL learning curves reflect reward expectations or whether they reflect learned policies (as predicted by Q learning vs actor-critic algorithms; Li and Daw, 2011; Jaskir and Frank, 2023). We thus sought to test these relationships directly by recording EEG during learning and then administering two retention tests. The EEG measures of RL were used to assess whether the neural RL measure is predictive of participants' ability to retrieve learned reward expectations and/or the retention of stimulus–response contingencies.

As a secondary aim, we also examined the impacts of acute stress on RL and WM processes. There is accumulating evidence, across various domains of learning, that acute stress reduces goal-directed decision-making and alters prefrontal cortex functioning (for review, see Arnsten, 2009), thereby promoting a shift from cognitively demanding but flexible systems toward simpler but more rigid systems (Kim et al., 2001; Schwabe and Wolf, 2009; Vogel et al., 2016; Wirz et al., 2018; Meier et al., 2022). We thus tested whether stress could reduce the ability of WM to effectively guide learning and instead enhance the relative contribution of RL processing.

Materials and Methods

Participants

Eighty-six healthy volunteers (43 women, age 18–34; mean = 24.56, SD = 3.84) participated in this experiment. All participants were right-handed, had normal or corrected-to-normal vision, and were screened for possible EEG contraindications. Individuals with a current medical condition, medication intake, or lifetime history of any neurologic or psychiatric disorders were excluded from participation. All participants provided written informed consent before the beginning of testing and received moderate monetary compensation. The study protocol was approved by the ethics committee of the Faculty of Psychology and Human Movement Sciences at the University of Hamburg.

Experimental procedure

Learning task. Interactions of RL and WM were tested using the RLWM task (Collins and Frank, 2012, 2018; Collins, 2018), programmed in MATLAB using the Psychophysics Toolbox. In this task (Fig. 1*A*), each trial started with a presentation of a stimulus in the center of the screen on a black background, and participants had to learn which of the three actions (key presses A1, A2, A3) to select based on trial-by-trial reward feedback. Stimulus presentation and response time were limited to 1.4 s. Incorrect choices led to feedback 0, whereas correct choices led to reward, (reward was 1 or 2 points fixed with the probability of 0.2, 0.5, or 0.8). Stimulus probability assignment was counterbalanced within participants to ensure equal overall value of different set sizes (see below) and motor actions. The key press was followed by audiovisual feedback (the word Win! with an ascending tone or the word Loss! with a descending tone). If participants did not respond within 1.4 s, the message Too slow! appeared. Feedback was presented for 0.4–0.8 s and was followed by a fixation cross for 0.4–0.8 before the next trial started.

To manipulate WM demands, the number of stimulus–action contingencies to be learned varied by block between one and five, denoted as not significant (ns), with a new stimuli set presented at each new block (e.g., colors, fruits, or animals). There were four blocks in which set size = 2, two blocks in which set size = 4, and three block in which set size = 1, 3, 5 for a total of 15 blocks and 645 trials. Within a block, each stimulus was presented 15 times; 108 stimuli were pseudorandomized, and 43 stimuli were presented for each participant. Stimulus category assignment to block set size was counterbalanced across subjects. Block order was also counterbalanced with the exception of set size = 1, which served as control (block numbers 8 and 14 were saved for set size = 1).

The following instructions were given to participants: In this experiment, you will see an image on the screen. You need to respond to each image by pressing one of the three buttons on the Gamepad: 1, 2, or three with your right hand. Your goal is to figure out which button makes you win for each image. You will have a few seconds to respond. Please respond to every image as quickly and accurately as possible. If you do not respond, the trial will be counted as a loss. If you select the correct button, you will gain points. You can gain either 1 or 2 points designated as "$" or "$$". Some images will give you more points for correct answers on average than other images. You can only gain points when you select the correct button for each image. At the beginning of each block, you will be shown the set of images for that block. Take some time to identify them correctly. Note the following important rules: There is ONLY ONE correct response for each image. One response button MAY be correct for multiple images, or not be correct for any image. Within each block, the correct response for each image will not change.

Test phase

After the learning phase, participants completed two surprise test phases (Fig. 1 *B*,*C*). The first was a reward retention test that has been used in earlier studies (Collins et al., 2017b). The reward retention test was designed to test whether expected values are learned by default as several previous studies showed that participants can select actions based on their relative expected values at the transfer phase even when they only had to learn which item was best (Frank et al., 2007; Palminteri et al., 2015). In this phase, on each trial participants were requested to select the more rewarding stimulus from a pair of stimuli that had each been encountered during the learning phase. All stimuli that were used in the

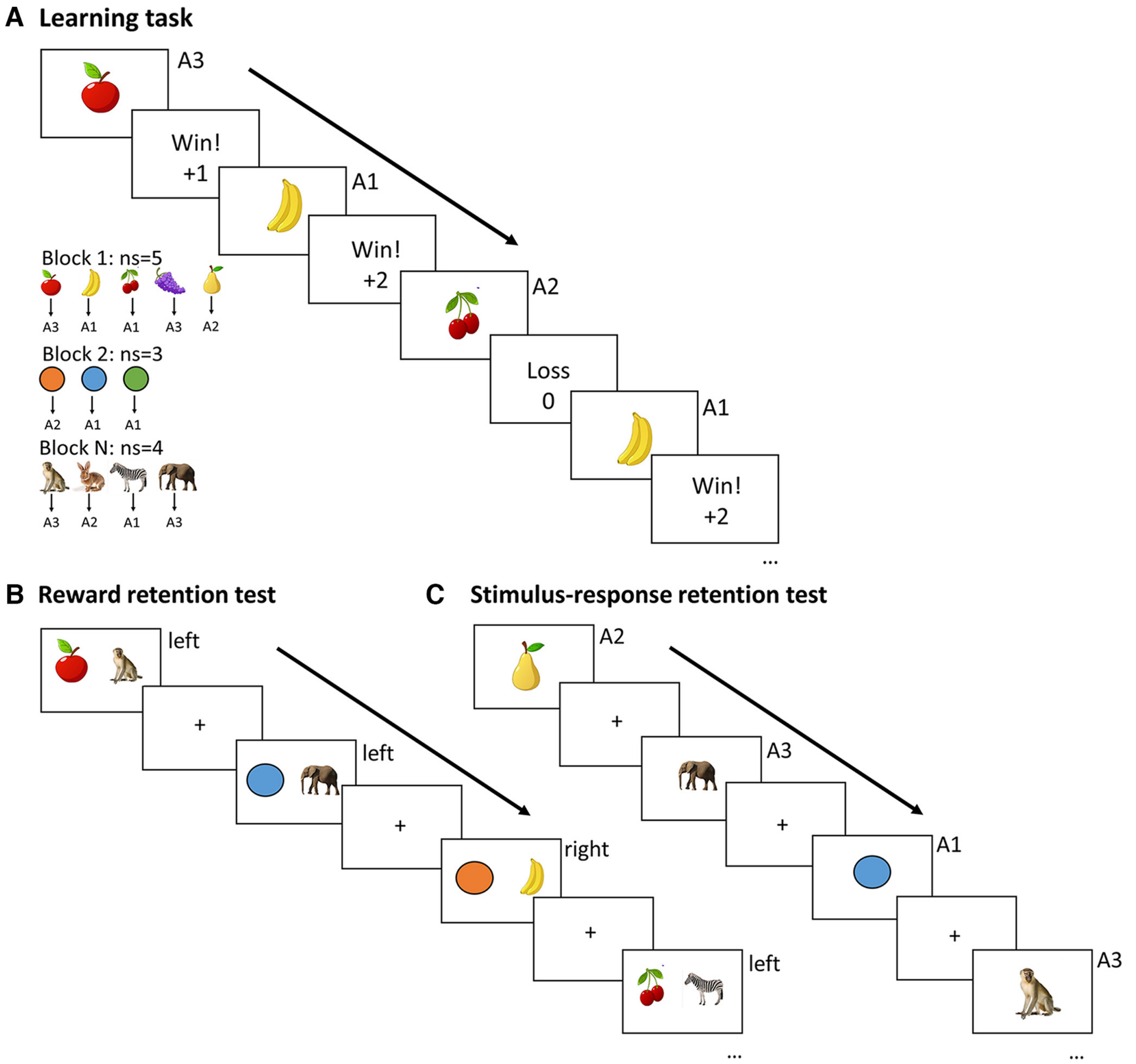


Figure 1. Experimental protocol of the learning task and the two test phases. ***A***, In the learning phase, in each block participants use deterministic reward feedback to learn which of three actions to select for each stimulus image. The set size (or the number of stimuli; ns) varies from one to five across blocks. After each response, feedback was presented audiovisually (see text for more details). ***B***, The surprise reward retention test protocol. In this task, participants are asked to recall the reward value of stimuli learned during the learning phase by choosing the stimulus they perceive to have been more rewarded within a pair of stimuli presented on every trial. ***C***, The surprise stimulus–response retention test protocol is a test of the learned stimulus–response policy. Here, participants are asked to recall the correct action for the probed stimulus. No feedback was given at either test phase.

learning phase were presented in the test phase at least once. The two stimuli were pseudorandomly selected to sample across all possible combinations of set sizes, blocks, and probabilities. To ensure no new learning at this phase, participants did not receive any feedback on their responses. Note that in this test, participants could not leverage information they had learned about which response to select (the policy); instead, they had to use novel response mappings to simply indicate which stimulus had been more rewarded. Participants' ability to select the more rewarding stimulus therefore required successful integration of the probabilistic reward magnitude history over learning for each stimulus.

The second test was the stimulus–response retention test, which assesses whether participants remember the correct response for each stimulus they had encountered previously during learning. Each of the stimuli used in the learning phase (except stimuli from block 1 and block 15 to limit primacy and recency effects) was presented four times individually, and participants were requested to press the key that was associated with the respective stimulus. Stimulus order was pseudorandomized to make sure that each stimulus was presented in each quarter of the test phase. No feedback was presented to rule out new learning during this test phase. Note that because this phase was preceded by the reward test phase, and because it followed many serial blocks of learning, it is not plausible that participants could hold information for previously encountered stimuli in WM, and thus retention depends on the memory for stimulus–action associations (the policy) as formalized by the RL system (Collins, 2018; Jaskir and Frank, 2023).

Behavioral data analysis

Statistical analyses were performed using R software (https://www.r-project.org/) and the lme4 package (version 1.1–26; Bates et al., 2015).

Data were fitted using generalized mixed-effect models (glmer) with the binomial family function. To avoid the Type I error rate without sacrificing statistical power, we followed the parsimonious mixed-model approach (Matuschek et al., 2017). We selected the random-effects structure that contained only variance components that were supported by the data by running singular value decomposition (Bates et al., 2015; Matuschek et al., 2017).

Behavioral analysis of learning task

To quantify the effect of RL versus WM, we analyzed learning performance (the proportion of correct responses) with general mixed-effect regression on trial-by-trial data from 86 participants as a function of both WM and RL variables and their interactions. The WM variables include the number of stimulus–response associations to be learned (denoted as *setSize*) and the number of intervening trials since the last time the stimulus was presented and a correct response was made (denoted as *delay*) reflecting WM interference or maintenance time in WM. The RL variable is the total number of previous correct (*Pcor*) responses for a stimulus. Participants and all the predictors were selected as random variables.

Behavioral analysis of the reward retention test

To quantify the possible effect of expected value learning under different WM loads, we analyzed test performance (the proportion of selecting the right vs left stimulus) with general mixed-effect regression on trial-by-trial data from 86 participants as a function of six variables, value difference (denoted as *delta_Q*, positive when the right stimulus had higher value and negative when the left stimulus had higher value); mean Q value of the stimulus pair [denoted as *mean value (Q)*]; mean set size of the stimulus pair (denoted as *mean_setSize*); the difference in set size (denoted as *delta_ setSize*, positive when the right stimulus was learned in higher set size); block (the block number in which they were learned, indicating how recently it was learned); and *perseveration* (binary coding of repetitions in response, repeat/switch). Participants, the effect of value difference (*delta_Q*), and the effect of set size difference (*delta_setSize*) were entered as random variables.

Behavioral analysis of the reward retention test together with EEG RL index

We ran a new regression model on the reward retention test data (including only the 77 participants that had EEG data), adding the difference in the EEG RL index between the pair of stimuli at choice. Because the neural RL index (see a detailed description of this measure below) could have both positive and negative values, all the predictors that were calculated as difference scores were taken as absolute scores and the model predicted performance accuracy (proportion of choosing the higher value stimulus). Test performance accuracy was analyzed as a function of the absolute model estimated value difference between the right and left stimulus (*abs_delta_Q*), the absolute difference in the EEG RL index between the right and left stimulus (*abs_delta_EEG_RL*), the mean value (estimated from the model) of the stimulus pair (*mean Q value*), the mean set size of the stimulus pair (*mean set size*), the absolute difference in the block number where the right and left stimulus were learned (*abs_ delta_block*), and response bias toward the previously selected response (*perseveration*; binary coding of repetitions in response). Participants, the effect of value difference (*abs_delta_Q*), and the effect of EEG RL index difference (*abs_delta_EEG_RL*) were entered as random variables.

Behavioral analysis of the stimulus–response retention test

In a general mixed-effect regression analysis, we tested accuracy for correctly recalling the response associated with a presented stimulus learned during the training phase as a function of set size (the set size block in which they were learned), block (the block number in which they were learned, indicating how recently it was learned), and model Q (the model estimated Q value of each stimulus calculated as the average Q value of the final six iterations during learning) and perseveration (the tendency to repeat the response selected in the previous trial at test coded as 1 for repeat and 0 for switch). The interactions between set size and model Q value, set size and block, and between set size and perseveration were also added as predictors. Participants and the interaction between model Q and set size were entered as random variables.

Behavioral analysis of the stimulus–response retention test together with EEG RL index

We ran the same regression model on the stimulus–response retention test data as before (including only the 77 participants that had EEG data), adding two new predictors, the average EEG RL index for each stimulus–response association (see a detailed description of this measure below) and the interaction between EEG RL index and set size. Participants, the interaction between model Q and set size, and the interaction between EEG RL index and set size were entered as random variables.

EEG recording and processing

During the learning task, participants were seated ~80 cm from the monitor in an electrically shielded and sound-attenuated cabin. EEG was recorded using a 64-channel BioSemi ActiveTwo system with sintered Ag/AgCl electrodes organized according to the 10–20 system. The sampling rate was 2048 Hz. The signal was digitized using a 24-bit A/D converter. Additional electrodes were placed at the left and right mastoids, ~1 cm above and below the orbital ridge of each eye and at the outer canthi of the eyes for measurement of eye movements. The EEG data were rereferenced off-line to a common average. Electrode impedances were kept below 30 kΩ. EEGs and EOGs were amplified with a low cutoff frequency of 0.53 Hz (= 0.3 s time constant).

The EEG data were processed using EEGLAB (Delorme and Makeig, 2004) and ERPLAB (Lopez-Calderon and Luck, 2014) toolboxes. The continuous EEG was bandpass filtered off-line between 0.5 and 20 Hz and downsampled to 125 Hz, then it was segmented into epochs ranging from 500 ms prestimulus up to 3000 ms poststimulus. The epoched data were visually inspected, and those containing large artifacts because of facial electromyographic activity or other artifacts, except for eyeblinks, were manually removed (e.g., large fluctuations in voltage across several electrodes that were in an order magnitude above neighboring activity). Independent components analysis was next conducted only on the 64 scalp electrodes using the EEGLAB runica algorithm. Components containing blink or oculomotor artifacts were subtracted from the data, resulting in an average of 1.6 components removed per participant (ranging between zero and three components). Finally, the epoched data were subjected to an automatic bad electrodes and artifact-detection algorithm (100 μV voltage threshold with a moving window width of 200 ms and a 100 ms window step), which was followed by manual verification. Bed electrodes were interpolated, and trials containing large artifacts were removed. Nine participants were removed from all the reported EEG analyses because of a high EEG artifact rate (>40% in one or more of the conditions) resulting in 77 participants who were used in the EEG analysis.

Data processing for behavior and EEG regression analysis

Omission trials, trials with very fast reaction times (RTs; <200 ms), and trials before the first correct response was made were excluded from all analyses. Setting the *delay* and *Pcor* variables to have one as their lowest level was done to ensure an interpretable analysis of these variables (Collins and Frank, 2012). The delay predictor (the number of trials since the stimulus was presented and a correct response was made) used in the regression analyses was inverse transformed (−1/delay) to avoid the disproportion effect of very large but rare delays (when a correct response was given early in the block but was then followed by several error responses for that stimulus).

Modeling

RL and WM contributions to participants' choices were estimated with the previously developed RLWM computational model (the model described below is identical to that used in Collins and Frank, 2018, where more details are provided). The RLWM is a mixture of a standard RL module with a delta rule and a WM module that has perfect memory for information that is within its limited capacity and is sensitive to delay (reflecting memory decay and interference from other intervening stimuli). For each stimulus–action association, the RL module estimates the

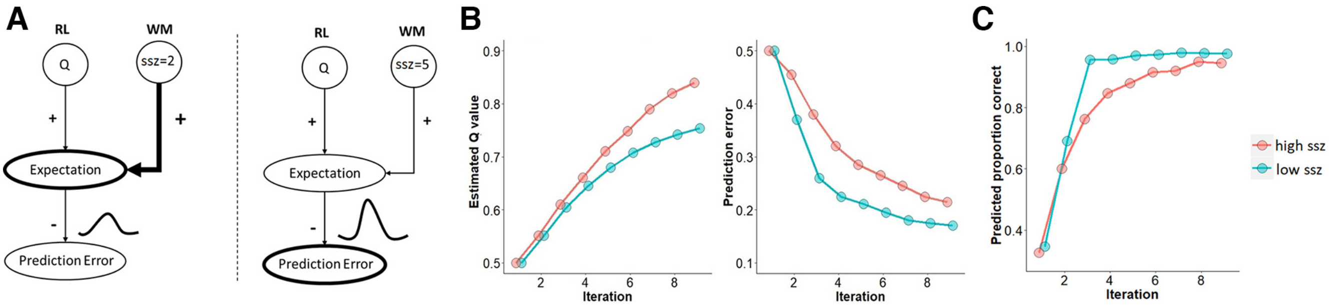


Figure 2. Cooperative interaction between the RL and WM systems (adapted from Collins and Frank, 2018). ***A***, Both WM and RL inform expected Q values and thus inform RPEs. When the number of stimuli to learn, set size (ssz) is within WM capacity (e.g., left, ssz = 2) the expected Q value of each contingency can be held in WM, thereby reducing RPEs during early learning compared with those that would occur from RL alone. When set size exceeds WM capacity (e.g., right, ssz = 5), degraded WM results in larger RPEs. ***B***, Computational model simulations (recreated from Collins and Frank, 2018) capture the RL and WM interaction, showing that larger RPEs persist for longer when WM load is taxed (high ssz), thereby accumulating expected Q values in the RL system. ***C***, Note that Q learning curves in ***B*** evolve more rapidly in high ssz, despite the opposite pattern in simulated behavioral learning curves (whereby WM contributes to rapid learning in low ssz).

expected value (Q) and updates those values incrementally on every trial as a function of the reinforcement history. This computation is complemented by the WM module, where information in the capacity-limited WM feeds into RL expectations, thereby affecting RL prediction errors and learning (Fig. 2).

Basic RL module. To maintain consistency with prior studies with this task and model, and to keep the model as simple as possible, we use Q learning for the model-free algorithm, but an actor-critic algorithm could also have been used (there are multiple options to capture incremental model-free RL, including methods that learn expected values for each choice and select on that basis; a canonical instance is Q learning and is often used in human studies) as well as methods that learn to directly optimize the policy (a canonical variant is an actor-critic model). Both classes of models similarly predict behavioral adjustment in RL tasks, and specific designs are needed to distinguish between them (Gold et al., 2012; Geana et al., 2022). The main goal here is to simply summarize the incremental RL process as distinct from the WM process.

Reward values were coded as zero or one for correct or incorrect (model fits are not improved if using one vs two points in the Q learning system, and behavioral learning curves are similar for stimuli that yield higher or lower probability of two points; Collins et al., 2017b). For each stimulus s and action a association, the RL module estimates the expected reward value Q and updates those values incrementally on every trial as follows:

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha \times \delta_t.$$

The Q value was updated as a function of the learning rate α (reflecting how fast reward expectations are updated) and the reward prediction error delta (δ), calculated as the difference between the observed reward R_t and the expected reward Q_t at each trial as follows: $\delta_t = R_t - Q_t$.

Choices were probabilistically determined using a softmax choice policy as follows:

$$p(a|s) = exp(\beta Q(s,a)) / \sum(exp(\beta Q(s,a_i))).$$

Here, β is the inverse temperature determining the degree to which differences in Q values are translated into more deterministic choices, and the sum is over the three possible actions. Q values were initialized to $1/n_A$, where $n_A = 3$ is the number of actions (i.e., the prior that any action is correct is one-third).

WM module. This module updates stimulus–action–outcome associations in a single trial. It assumes that stimulus–action–outcome information, when encoded and maintained in WM, could serve to update reward expectation rapidly and accurately (i.e., perfect retention of information from the previous trial). When not limited by capacity and decay (see below), the WM module is therefore represented by a Q learning system with a learning rate of 1 ($\alpha = 1$).

Decay. To account for potential forgetting on each trial because of delay or WM interference, we included a decay parameter ϕ $(0 < \phi < 1)$, which pulls the estimates of Q values toward their initial value [$Q_0 = 1/n_A$, number of actions $n_A = 3$] as follows:

$$Q \leftarrow Q + \phi(Q_0 - Q).$$

Only the WM module was subject to forgetting (decay parameter φ_{WM}) to capture the well-documented short-term stability of WM in contrast to the robustness of RL.

WM contributes to choice. Because WM is capacity limited, only K stimulus and action associations can be remembered. A constraint factor reflects the a priori probability that the item was stored in WM as follows: $w_{WM}(0) = P_0\ (WM) = K/n_s$ (i.e., the set size in the current block relative to capacity K) and implies that the maximal use of WM policy relative to RL policy depends on the probability that an item is stored in WM. This probability is then scaled by ρ $(0 < \rho < 1)$, the participant's overall reliance of WM versus RL (where higher values reflect greater confidence in WM), in the following:

$$w_{WM}(0) = \rho * min(1, K/n_s).$$

Cooperative model. Although the original model (Collins and Frank, 2012) assumed independent RL and WM modules that compete to guide behavior, our more recent work suggests that WM expectations influence RL updating (Collins and Frank, 2018). Thus, WM contributes part of the reward expectation for the RL model, according to the following equation: $\delta_t = R_t - [w_{WM} \times Q_{WM} + (1 - w_{WM}) \times Q_{RL}]$, where w_{WM} is the weighting parameter (the degree to which WM is weighted relative to RL, which is stronger in low set sizes), and Q_{WM} is the expected reward from the WM module. This RPE is then used to update the RL Q value as follows: $Q_{t+1} = Q_t + \alpha \times \delta_t$.

This interactive computation of RL forms the basis of the simulated predictions shown in Figure 2. Nevertheless, as explained in Collins and Frank (2018), we test these predictions by fitting models in which RL and WM modules are independent. (Independence is assumed in the original models, which still provide good fits to the data because when information is within WM, WM dominates updating and contributes to rapid learning curves, and hence the smaller RPEs and RL Q values of the interactive models for small set sizes are not influential on behavioral accuracy during learning; however, this model makes differential predictions for neural learning curves and future retention.) We then assess systematic deviations from independence informed by these simulations (e.g., neural Q learning curves should grow more rapidly in high than in low set sizes; Fig. 2).

Data processing for univariate EEG analysis

To extract the neural correlates in the EEG signal of conditions of interest, we used a mass univariate approach (Collins and Frank, 2018). A

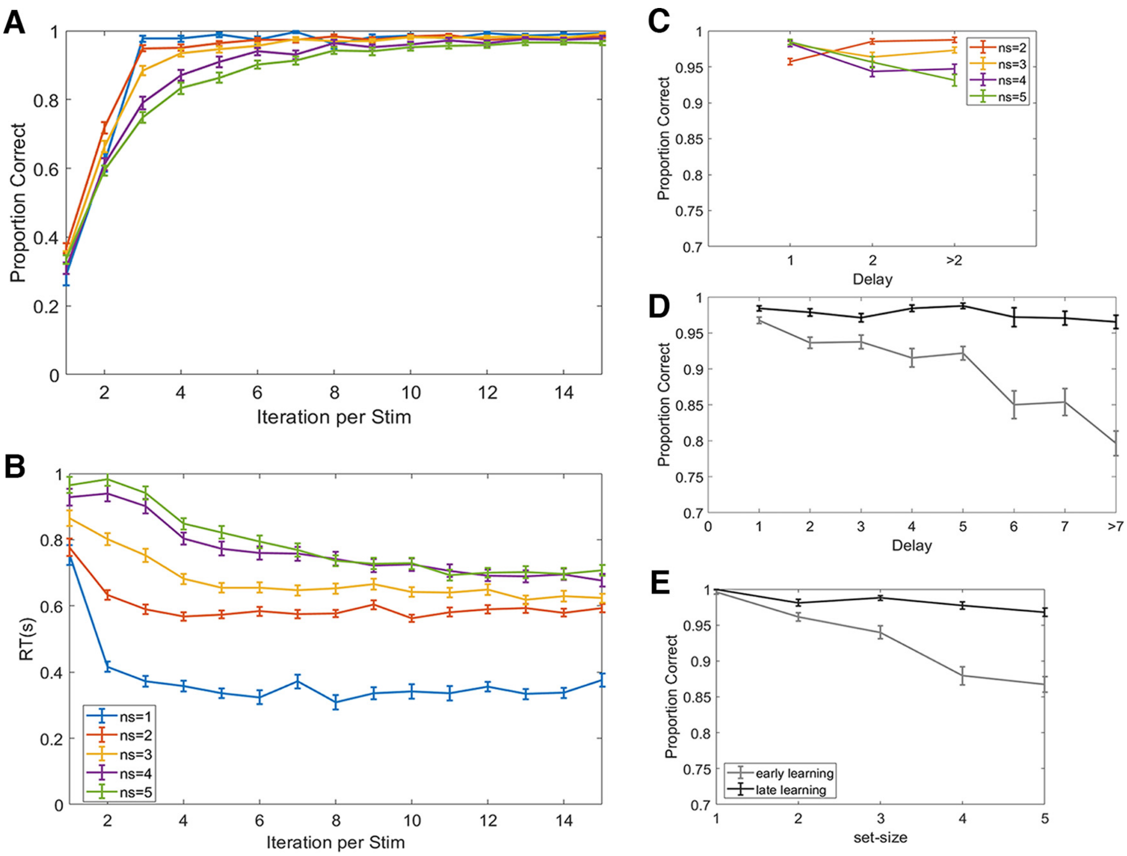


Figure 3. Behavioral results from the learning phase. ***A–B***, Performance learning curves and RTs for each set size as a function of the number of iterations of a stimulus (stim). ***C***, Performance as a function of WM load, the detrimental effect of delay is greater in high set sizes. ***D–E***, Reduced effects of both delay and set size as learning progresses from early (up to 2 previous correct choices) to late (the last 2 trials of each stimulus) trials in a block, suggestive of a transition from WM to RL.

multiple regression analysis was conducted for each participant in which the EEG amplitude at each electrode site and time point was predicted by the conditions of interest, the set-size (number of stimulus–response–outcome associations given in a block), model-derived RL expected value (denoted as Q), delay (number of trials since this stimulus was presented and a correct response was given), and the interaction of these three regressors while controlling for other factors like reaction time (log transformed) and trial number within block. Furthermore, the EEG signal was reduced to a selected window of −100 to +700 ms around stimulus onset and was baseline corrected from −100 to 0 ms before the onset of the stimulus. To account for remaining noise in the EEG data, the EEG signal (at each time point and electrode) was *z*-scored across all trials and so were all the predictors before they were entered to the robust multilinear regression analysis (Collins and Frank, 2018).

Corrected ERPs

To plot corrected ERPs, we computed the predicted voltage using the multiple regression model described above while setting a single regressor to zero (set size, delay, expected Q value, or reaction time); we subtracted this predicted voltage from the true voltage (for every electrode and time point within each trial), leaving only the fixed effect, the variance explained by that regressor, and the residual noise of the regression model. ERPs were computed as the average corrected voltage from all trials that belong to the same level of condition. Note that the array of expected Q values was divided to four quartiles, and trials within each quartile were averaged for plotting ERPs.

Trial-by-trial similarity index of WM and RL

As explained above, a multiple regression analysis was conducted for each participant in which the EEG amplitude at each electrode site and time point was predicted by the conditions of interest (set size, delay, RL expected value, and their interactions). We used the previously identified analysis method (Collins and Frank, 2018; Rac-Lubashevsky and Frank, 2021) to identify spatiotemporal clusters (masks) of the three main predictors in the GLM (set-size, delay, and model-derived RL expected value). Specifically, we tested the significance of each time point at each electrode across participants against zero using only trials with correct responses.

We then used cluster-mass correction by permutation testing with custom-written MATLAB scripts. Cluster-based test statistics were calculated by taking the sum of the *t* values within a spatiotemporal cluster of points that exceeded the $p = 0.001$ threshold for a *t* test significance level. This was repeated 1000 times, generating a distribution of maximum cluster-mass statistics under the null hypothesis. Only clusters with a greater *t* value sum than the maximum cluster mass obtained with 95% chance permutations were considered significant. We then assessed the neural similarity of each trial to the spatiotemporal mask by computing the dot product between the activity in the individual trial (voltage maps of electrode × time) and the identified masks (*t* value maps of electrode × time). This computation produced a trial-level similarity measure intended to assess the trial-wise experienced WM load and delay effects, as well as trial-wise RL contributions.

The EEG RL index predictor used in the general mixed-effect regression analyses of both test phases was calculated by averaging the EEG RL index in the final six iterations of each stimulus. This was done for each stimulus–response association within each participant.

Stress manipulation

All testing took place in the morning between 8:00 A.M. and noon. On their arrival in the lab, participants' baseline measures of blood pressure and salivary cortisol were taken. Afterward, participants were prepared for the EEG and completed the Multidimensional Mood State Questionnaire (Steyer et al., 1994) that measures subjective mood on the scales, negative versus elevated mood, calmness versus restlessness,

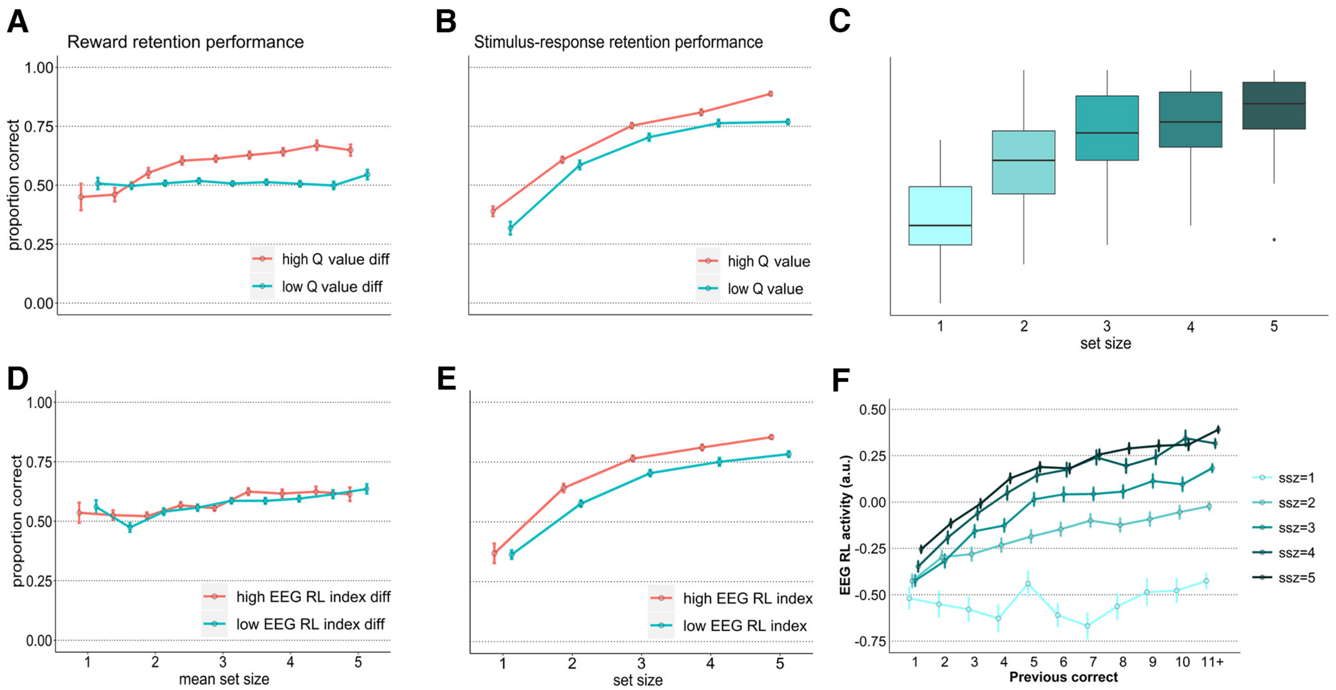


Figure 4. Behavior performance at the test phase. ***A***, Effect of value difference and set size on the reward retention test performance. The proportion of correct selection of the more rewarding stimulus from a pair of the probed stimuli increases as a function of differences in the number of experienced rewards (Q value diff) and the set size in which they were learned. diff, Difference. The median split of absolute value differences is shown (red, high-Q value difference trials; blue, low-Q value difference trials). ***B–C***, Effect of set size on the stimulus–response retention test performance. The proportion of correct recall in the test phase increases as a function of the estimated Q values of the probed association and as a function of the set size in which it was learned. The median split of the estimated stimulus–response Q values is shown (red, high Q value associations; blue, low Q value associations). ***D***, Effect of EEG RL index on the reward retention test performance. The proportion of correct selection of the more rewarding stimulus from a pair of the probed stimuli increases as a function of the set size in which they were learned but was not further modulated by the magnitude of the EEG RL index of the stimuli. The median split of absolute differences in EEG RL indices is shown (red, high-EEG RL index difference; blue, low-EEG RL index difference). ***E***, Effect of the neural RL index on recall accuracy in the stimulus–response retention test. The neural RL index is shown as the median split across all the RL indices. Stimuli with high RL index are depicted in red and stimuli with low RL index are depicted in blue. ***F***, The EEG RL index increases parametrically with the increase in accumulated rewards. These neural learning curves parametrically increase with set size. Error bars indicate SE.

and wakefulness versus tiredness, before and after the treatment as well as after the learning task. Forty-two participants underwent the Socially Evaluated Cold Pressor Test (SECPT; Schwabe et al., 2008), and 44 participants were assigned the warm water control condition. The SECPT is a standardized stress protocol in experimental stress research that combines physiological and psychosocial stress elements and has been shown to result in robust stress responses (Schwabe and Schächinger, 2018). During the SECPT, participants in the stress group immersed their right hand for 3 min in ice water (0–2°C) while being videotaped and evaluated by a nonreinforcing, cold experimenter. In the control condition, participants immersed their hands in warm water (35–37°C), without being videotaped or evaluated by an experimenter. About 25 min after the treatment, participants received the learning task instructions and completed a brief training session, after which they completed the learning task and test phases 1 and 2. In total, the experiment lasted ~130 min.

Results

In line with previous findings in this task (Collins et al., 2017b), our data demonstrated separable contributions of RL and WM systems to performance. The contribution of incremental RL was observed as the proportion of correct responses increased with the progress in the block (Fig. 3*A*) and with the increase in reward history [*Pcor*, $\beta = 0.67$, SE = 0.05, $z(46926) = 13.17$, $p < 0.001$]. WM contributions were observed as learning was strongly affected by set size with a greater proportion of correct responses in low set sizes than in high set sizes [set size, $\beta = -0.28$, SE = 0.05, $z(46926) = -5.39$, $p < 0.001$]. Learning curves were more gradual in higher set sizes than in low set sizes (Fig. 3*A*; and slower, Fig. 3*B*). Moreover, performance decreased with increasing delay in larger set sizes [*delay* × ns, $\beta = -0.09$, SE = 0.05, z(46926) = −2.59, $p = 0.009$; Fig. 3*C*]. These relative contributions of WM decreased with learning as the detrimental effect of delay attenuated with the increase of accumulated rewards [ns × *Pcor*, $\beta = 0.13$, SE = 0.04, z(46926) = 3.35, $p < 0.001$; *delay* × *Pcor*, $\beta = 0.34$, SE = 0.04, z(46926) = 9.17, $p < 0.001$; ns × *delay* × *Pcor*, $\beta = 0.20$, SE = 0.03, $z(46\,926) = 6.37$, $p < 0.001$; Fig. 3*D*,*E*], reflecting a transition from WM to RL. Together these results confirm the cooperative interaction of early WM contributions that diminish as RL becomes more dominant.

Behavioral performance: reward retention test

Results replicated previous findings in this phase (Collins et al., 2017b). Participants were more likely to select the stimulus for which they had been rewarded more often during learning as a function of the difference between the number of rewards experienced for these stimuli [*delta_Q*, $\beta = 0.41$, SE = 0.04, $z(19796) = 9.76$, $p < 0.001$]. Moreover, also replicating previous findings, this value discrimination effect was enhanced when stimulus values were learned under higher set sizes rather than under lower set sizes [*mean_setSize* × *delta_Q*, $\beta = 0.11$, SE = 0.02, $z(19796) = 6.04$, $p < 0.001$]. For display purposes, the median split in the absolute delta_Q score is shown as high- and low-value differences (Fig. 4*A*). Furthermore, participants were generally less likely to select the stimulus learned under a higher set size than under a low set size [*delta_setSize*, $\beta = -0.69$, SE = 0.09, $z(19796) = -7.61$, $p < 0.001$], an effect previously attributed to participants learning a cost of mental effort in a high set size (Collins et al.,

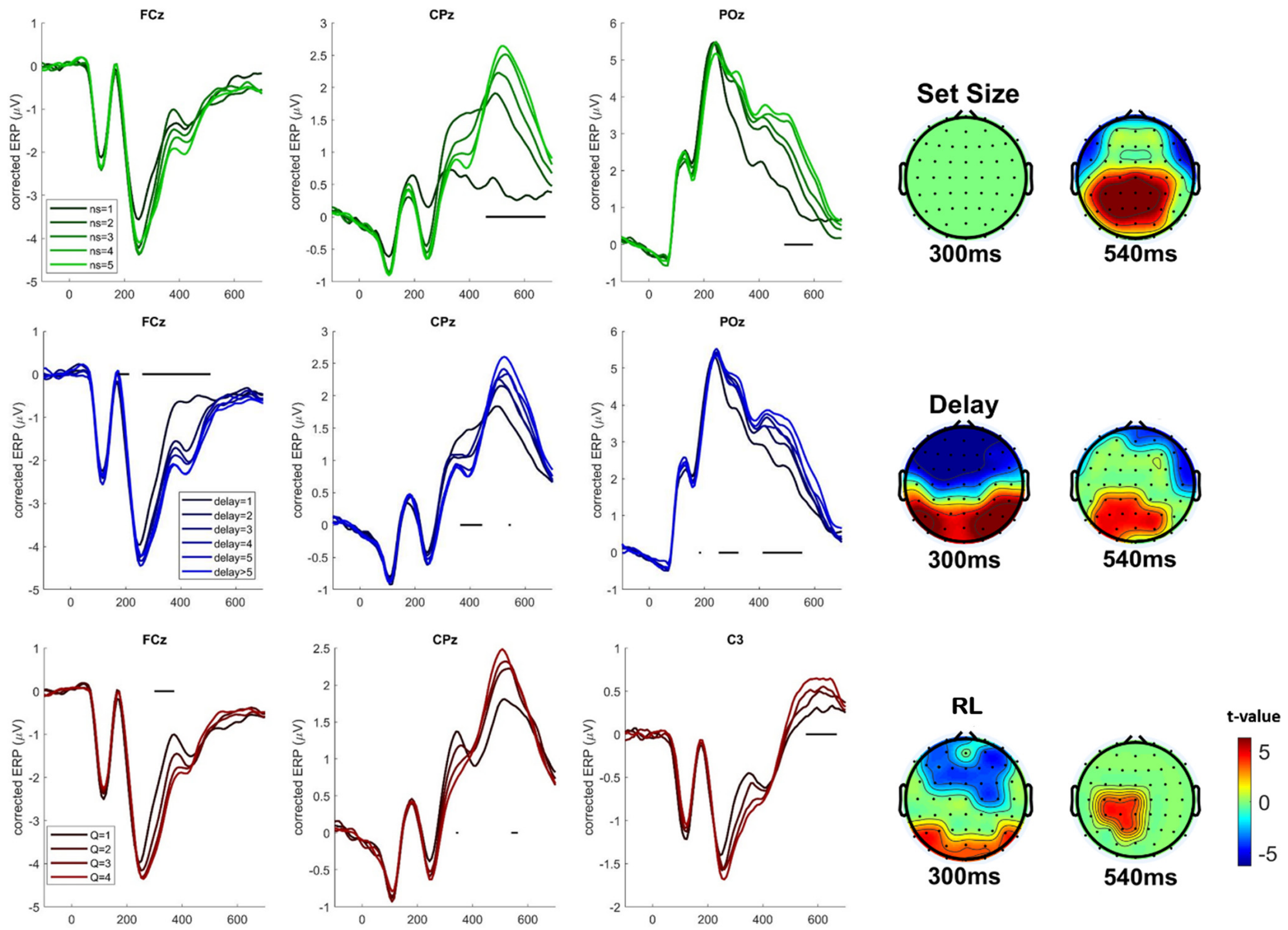


Figure 5. EEG decoding of RL and WM effects during choice. Corrected ERPs exhibiting the effect of three main predictors (top to bottom rows; green, set size; blue, delay; red, RL value quartiles) on the voltage of significant electrodes (FCz, CPz, and Poz for set size and delay, and FCz, CPz, and C3 for RL). The black line reflects the significant time points after permutation correction. Right, The effect of each predictor in the row is exhibited with a scalp map topography at early (300 ms) and late (540 ms) time points. The color in the scalp map represents significant thresholded *t* values.

2017b). There was no effect for the difference in the block in which the item values were learned, nor was the set size effect modulated by block number ($p > 0.82$). We also controlled for response perseveration; no significant tendency was observed for repeating the same response used in the previous trial ($p > 0.69$).

Behavioral performance: stimulus–response retention test

Supporting the key model prediction that retention of stimulus–response associations should improve as load increases, we observed better recall performance for associations learned under high rather than low set sizes [set size, β = 0.84, SE = 0.05, $z(11894) = 15.83$, $p < 0.001$]. And, indeed, this effect was parametric, with substantially better performance as set size increased (Fig. 4*B*,*C*). This effect is particularly striking given that performance is parametrically worse for the higher set size items during learning (compare Fig. 3*A*, Fig. 4*C*). Not surprisingly, recall accuracy in the test phase was positively predicted by the estimated Q value of the probed stimulus–response association [*model Q*, β = 0.27, SE = 0.04, $z(11\,894) = 6.97$, $p < 0.001$]; that is, associations that were learned better were also better remembered. Importantly, this effect grew when the set size was high [*model Q* × set size, β = 0.15, SE = 0.04, $z(11\,894) = 3.64$, $p < 0.001$; Fig. 4*B*]. Recall accuracy was also subject to the influence of recency as associations learned during more recent than early blocks were also recalled more accurately [block, β = 0.22, SE = 0.03, $z(11894) = 8.61$, $p < 0.001$]. This recency effect increased for associations learned under higher set sizes [set size × block, β = 0.09, SE = 0.02, $z(11894) = 4.13$, $p < 0.001$]. No effect of perseveration in responses was observed ($p > 0.11$).

EEG correlates of WM and RL during learning

The model-based EEG analysis indicated significant effects for all three variables of interest—set size, delay, and RL. Consistent with previous EEG results in this task (Collins and Frank, 2018) and with the prediction that separable systems contribute to learning, the neural signals of RL exhibited an early frontal activity (~300 ms poststimulus onset; Fig. 5) that preceded the parietal neural signal of set size (peaked at ~540 ms; Fig. 5), supporting the engagement of the RL system early in the trial followed by the cognitively effortful WM process. The neural signals of RL exhibited an additional late temporal activity (~600 ms poststimulus onset) that overlapped in time with the set size effect. Finally, a significant frontal and parietal effect of delay was also observed to initiate early at 300 ms.

To quantify how the neural measure of RL is modulated by WM and RL processes, we analyzed the trial-by-trial level EEG RL index (reflecting how strong the RL computation is at a given trial) with linear effects regression from 77 participants, as a function of set size (*setSize* = 1, 2, 3, 4, 5), the number of previous correct (*Pcor* = 1:15), and the interactions between them (see above, Materials and Methods). As expected because of incremental learning, neural indices of RL increased parametrically as a function of reward history (*Pcor*, $\beta = 0.17$, $t_{(38,377)} = 34.77$, $p < 0.001$). Importantly, confirming model predictions, neural RL signals increased to a larger extent as the set size grew (*Pcor* × *setSize*, $\beta = 0.04$, $t_{(38,377)} = 7.53$, $p < 0.001$; Fig. 4*F*). This finding corroborates previous reports that RL computations are larger in high set sizes because of diminishing WM contributions and thus increasing the accumulation of reward prediction errors (Collins et al., 2017b; Collins and Frank, 2018).

We next assessed the core prediction that the neural RL index is related to future retention, and more specifically the cooperative model prediction that the speeded neural RL curves in high set sizes are related to better retention of learned contingencies. Notably, although this prediction did not hold for the reward retention phase (*abs_delta_EEG_RL*, $p = 0.65$; *mean_setSize* × *abs_delta_EEG_RL*, $p = 0.61$; Fig. 4*D*), it was clearly borne out for the stimulus–response retention phase [EEG RL, $\beta = 0.23$, $z(10613) = 4.51$, $p < 0.001$; Fig. 4*E*]. Stimuli that had been associated with a larger EEG RL index during learning were associated with better recall of the associated response at test; this effect held even when controlling for the non-neural predictors (which replicated the prior analysis). Figure 4*E* shows that a high EEG RL index (by median split) was predictive of better retention performance at test. The finding that the neural index of RL is related to policy retention but not reward retention is relevant for models that dissociate whether model-free RL in the brain encodes expected values or policies (see above, Materials and Methods, model method; see below, Discussion). Note that a slightly different regression model was used for testing the neural RL index effect on the reward retention test performance from the behavior model used previously (see above, Materials and Methods). Nevertheless, the key behavior results were replicated in this analysis as performance increased with the increase in the absolute value differences [*abs_delta_Q*, $\beta = 0.31$, SE = 0.03, $z(17743) = 8.82$, $p < 0.001$], and although this effect was not further modulated by set size (*mean_setSize* × *abs_delta_Q*, $p = 0.63$), performance accuracy did improve with set size [*mean_setSize*, $\beta = 0.07$, SE = 0.02, $z(17743) = 3.23$, $p = 0.001$; Fig. 4*D*].

Acute stress modulation of RL and WM interaction

Manipulation check

Subjective, autonomic, and endocrine data indicated that the stress induction by the SECPT was successful. The SECPT was rated as significantly more unpleasant, stressful, and painful than the warm water control procedure (more difficult, $t_{(84)} = 9.941$, $p < 0.001$, $d = 2.14$; more unpleasant, $t_{(84)} = 9.088$, $p < 0.001$, $d = 1.96$; more stressful, $t_{(84)} = 7.72$, $p < 0.001$, $d = 1.66$; and more painful $t_{(84)} = 11.42$, $p < 0.001$, $d = 2.46$; Table 1 and Table 2). Furthermore, we observed significant Treatment-by-Time interactions for subjective stress ratings (negative mood, $F_{(2,164)} = 10.53$, $p < 0.001$, $\eta^2_g = 0.02$; restlessness, $F_{(2,164)} = 9.47$, $p < 0.001$, $\eta^2_g = 0.02$) and autonomic arousal measures (systolic blood pressure, $F_{(4,336)} = 26.22$, $p < 0.001$, $\eta^2_g = 0.06$; diastolic blood pressure, $F_{(4,336)} = 26.99$, $p < 0.001$, $\eta^2_g = 0.09$; and heart rate, $F_{(3,252)} = 10.70$, $p < 0.001$, $\eta^2_g = 0.02$). As expected, these autonomic responses returned relatively quickly to baseline after the treatment (Fig. 6). The stress and no-stress control groups did not differ in any of the autonomic arousal measures pretreatment (all p values > 0.07).

Table 1. The mean and SD (in parentheses) of the ratings before and after the procedures are reported for the control group

Control group			
Procedure ratings	Before	After	End of testing day
Subjective mood			
Depressed mood vs elevated mood	33.69 (4.99)	34.26 (4.72)	33.86 (4.66)
Restlessness vs calmness	32.476 (6.08)	33.83 (5.14)	33.24 (4.61)
Sleepiness vs wakefulness	28.571 (6.48)	28.31 (6.88)	26.64 (6.78)
Rating of control procedure			
Difficult	—	4.09 (13.21)	—
Unpleasant	—	9.52 (21.88)	—
Stressful	—	4.20 (15.23)	—
Painful	—	3.79 (14.62)	—

Table 2. The mean and SD (in parentheses) of the ratings before and after the procedures are reported for the stress group

Stress group			
Procedure ratings	Before	After	End of testing day
Subjective mood			
Depressed mood vs elevated mood	33.76 (3.51)	31.57 (5.32)	33.43 (3.99)
Restlessness vs calmness	32.99 (4.24)	30.45 (6.14)	32.43 (4.72)
Sleepiness vs wakefulness	28.98 (5.71)	29.86 (6.16)	26.45 (6.12)
Rating of stressor			
Difficult	—	50.69 (28.01)	—
Unpleasant	—	58.73 (28.09)	—
Stressful	—	40.17 (26.70)	—
Painful	—	55.40 (25.97)	—

Salivary cortisol (sCORT) responses were assessed by running ANOVA with Time (T1, T2, T3, T4) as the within-subject factor and Treatment (SECPT vs warm water control group) as the between-subject factor. We observed a significant effect for Time ($F_{(3,234)} = 28.53$, $p < 0.001$, $\eta^2_p = 0.27$) but not for Treatment ($F_{(1,78)} = 3.03$, $p = 0.08$, $\eta^2_p = 0.04$). An expected Treatment × Time interaction was observed ($F_{(3,234)} = 6.97$, $p < 0.001$, $\eta^2_p = 0.08$), with the stress group displaying greater sCORT levels immediately before the learning task (23 min posttreatment; $t_{(78)} = 2.80$, $p = 0.006$, $d = 0.63$), but only marginal difference was observed at half time during learning task (50 min post-treatment; $t_{(78)} = 1.90$, $p = 0.06$, $d = 0.43$). No difference in sCORT levels was observed at baseline ($t_{(78)} = 0.61$, $p = 0.54$), nor at the end of the learning task (80 min posttreatment; $t_{(78)} = 0.11$, $p = 0.91$), suggesting that stress-induced cortisol elevations gradually decreased during the learning task (Fig. 6). Note that six participants were excluded from the cortisol analysis because they did not provide sufficient saliva for analysis.

Learning phase performance by stress group

To test the hypothesis that acute stress may reduce the ability of WM to effectively guide learning, thereby weakening the relative contribution of WM in the training phase in the stress group compared with the control group, we ran the same general mixed-effect regression model on trial-by-trial training data from 86 participants but added stress group as a factor (42 participants in the stress group and 44 participants in the control group). This analysis revealed that learning by set size interaction was modulated by stress [*Pcor* × set size × stress_group, $\beta = -0.20$, SE = 0.08, $z(46926) = -2.60$, $p = 0.009$] and so was the learning by delay interaction [*Pcor* × *delay* × stress_group, $\beta = 0.22$, SE = 0.07, $z(46\,926) = 3.04$, $p = 0.002$]. To understand the nature of these interactions, we ran two follow-up analyses using

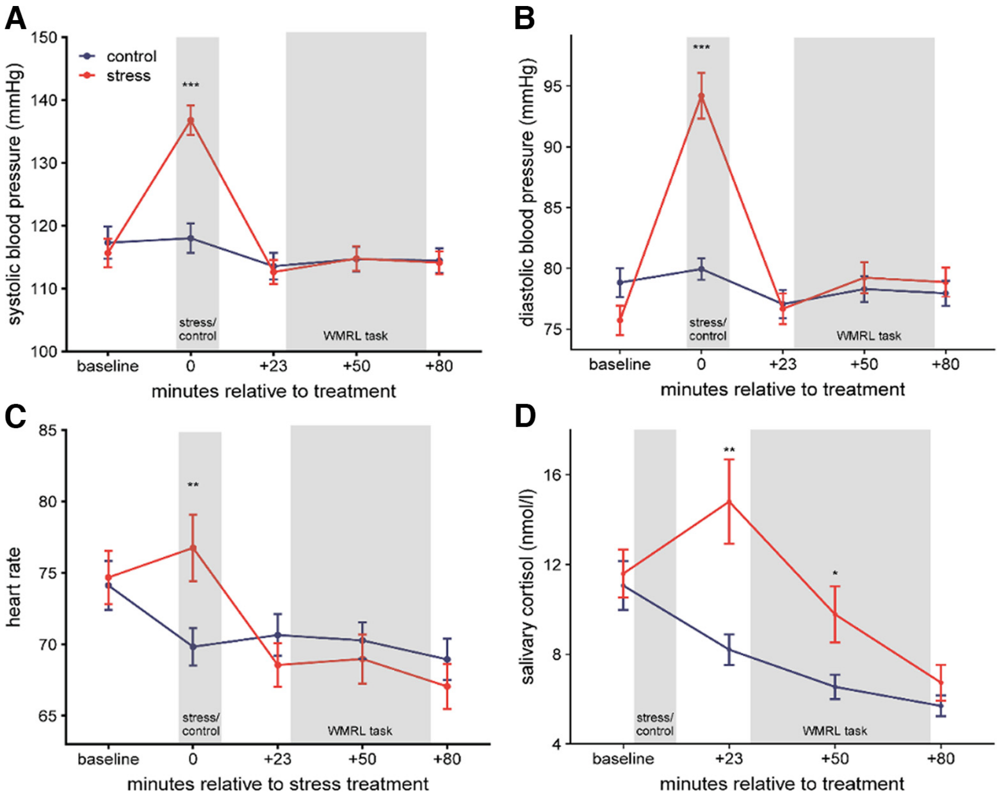


Figure 6. Successful stress induction. ***A–D***, The exposure to the stressor led to significant increases in systolic blood pressure (***A***), diastolic blood pressure (***B***), heart rate (***C***), and salivary cortisol levels (***D***). Error bars indicate SEs. The control group is depicted in dark blue and the stress group in red; $**p < 0.01$, $***p < 0.001$ for the comparison between the stress group and the control group.

the same general mixed-effect regression model on trial-by-trial training data, separately in the control ($N = 44$) and the stress group ($N = 42$). These analyses showed that learning curves were additive to the set size effect in the stress group (*Pcor* × set size, $p = 0.74$) but not in the control group [*Pcor* × set size, $\beta = 0.22$, SE = 0.05, $z(24\,031) = 4.30$, $p < 0.001$], which showed a greater drop in performance during high set sizes (Fig. 7*A*,*B*). The attenuated delay effect with learning was significant for both the stress group [*Pcor* × *delay*, $\beta = 0.47$, SE = 0.05, $z(22895) = 8.41$, $p < 0.001$] and the control group [*Pcor* × *delay*, $\beta = 0.23$, SE = 0.05, $z(24031) = 4.74$, $p < 0.001$; Fig. 7*C*,*D*].

Reward retention test performance by stress group

To test the hypothesis that acute stress may reduce the ability of WM to effectively guide learning, thereby strengthening RL conurbations during the training phase and leading to better retention of learned information in the stress group compared with the control group, we ran the same general mixed-effect regression model on trial-by-trial reward retention test data from 86 participants but added stress group as a factor (42 participants in the stress group and 44 participants in the control group) and analyzed test performance (the proportion of selecting the right vs left stimulus). This analysis replicated the results of the behavior analysis without the group factor. No effect of stress was observed ($p > 0.15$; Fig. 7*E*).

Stimulus–response retention test performance by stress group

To test the hypothesis that acute stress may reduce the ability of WM to effectively guide learning, thereby strengthening RL conurbations during the training phase and leading to better retention of learned information in the stress group compared with the control group, we ran the same general mixed-effect regression model on trial-by-trial stimulus–response retention test data from 86 participants but added stress group as a factor (42 participants in the stress group and 44 participants in the control group) and analyzed test performance. This analysis revealed that the effect of set size on recall accuracy of stimulus–response associations interacted with stress [set size × stress_group, $\beta = 0.22$, SE = 0.10, $z(11894) = 2.30$, $p = 0.02$; Fig. 7*F*], but follow-up analysis on each group separately showed significant effect of set size on recall accuracy in both the control group [$\beta = 0.72$, SE = 0.07, $z(6129)=10.72$, $p < 0.001$] and the stress group [$\beta = 0.95$, SE = 0.08, $z(5765) = 11.76$, $p < 0.001$].

Discussion

Together, our findings provide insight into the intricate interplay between WM and RL during learning, and its opposing influences on acquisition versus retention of stimulus–response associations. A previous study proposed a cooperative WMRL model, whereby RPEs in the RL system are not only computed relative to RL expected values but are also modulated by expectations held in WM (Collins and Frank, 2018). This model accounted for fMRI and EEG findings in which neural RPEs were diminished for smaller WM loads (Collins et al., 2017a; Collins and Frank, 2018). Moreover, this model accounted for findings that on a given trial, larger neural indices of WM expectations were predictive of subsequent RPEs during the outcome, even within a given set size (Collins and Frank, 2018). This model led to a key prediction that enhanced RL processes under high WM load would support more robust retention of learned association, despite the substantially slower acquisition. Preliminary behavioral

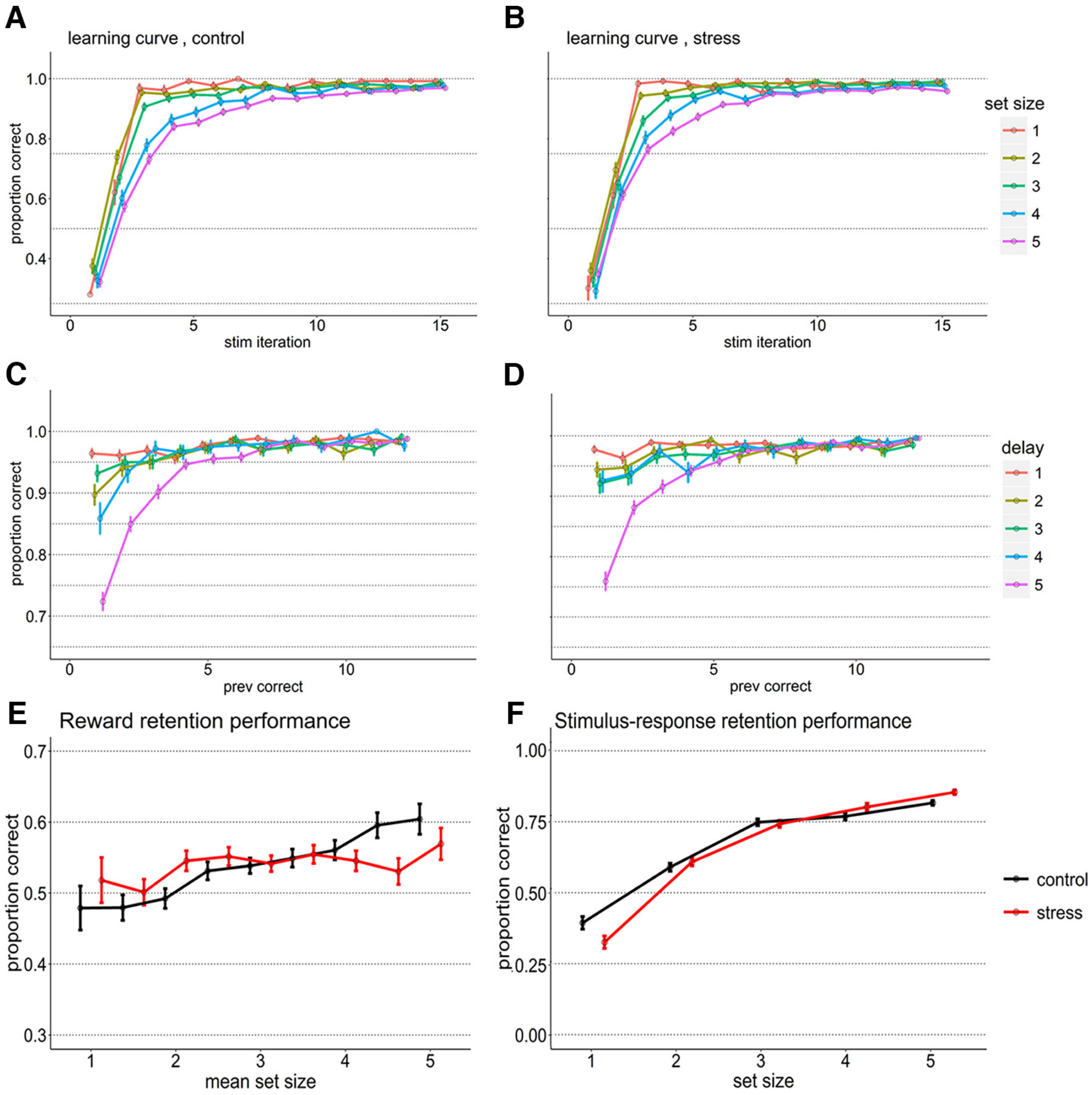


Figure 7. Stress effects during the learning and test phases. ***A***, ***B***, Learning curves across iterations as a function of set size in the control group (***A***) and stress group (***B***). ***C***, ***D***, Learning curves across the number of previous correct as a function of delay (1–5 where 5 reflects delay of 5 and above) in the control group (***C***) and stress group (***D***). ***E***, Effect of stress on the reward retention test performance. The proportion of correct selection of the more rewarding stimulus from a pair of the probed stimuli increases as a function of the set size in both the control group (black) and in the stress group (red). ***F***, Effect of stress on recall accuracy in the stimulus–response retention test. The proportion of correct recall in the stimulus–response test increases as a function of the set size in both the control group (black) and the stress group (red). Error bars indicate SEs.

evidence for such a behavioral prediction had been reported by Collins (2018), who showed enhanced retention of items learned in set size 6 compared with set size 3. However, that study did not employ neural recordings and thus did not test whether the neural WMRL interaction was the underlying mechanism for these effects. Here, we provide several lines of evidence in support of this claim.

First, our behavioral and EEG results replicated key findings in the RLWM task and in the subsequent memory tests. In the learning task, we observed worse acquisition with increasing set size and with delays between successive stimulus presentations, but as learning progressed (with the increase in reward history) the negative effect of delay in high set sizes diminished considerably. This observation further supports the model prediction that RL dominates over WM with the accumulation of rewards over time. Second, at the neural level, we also replicated findings in which neural RL indices preceded the cognitively costly WM process during stimulus processing (Collins and Frank, 2018). Moreover, we found robust evidence that EEG signals of RL increased more rapidly across trials under high than low load (Fig. 4*F*), a key prediction of the cooperative model (Fig. 2), although behavioral learning was slower in these conditions.

Importantly, we observed that associations learned under higher WM load had increasingly higher recall accuracy in the stimulus–response retention test (Fig. 4*C*). This result extends the previously reported retention benefit of associations learned under high compared with low set sizes (Collins, 2018). We showed that this effect is parametric across five levels of WM load, and moreover that the greatest retention deficits occurred for the very lowest set sizes in which participants could easily learn the task purely via WM. Furthermore, we replicated previous results in the reward retention test (Collins et al., 2017b) and demonstrated that participants have differential sensitivity to the proportion of trials in which they were rewarded for either of the stimuli and this effect grew with set size.

Finally, to gain a better understanding of the mechanism responsible for the benefits in both retention tests, we leveraged a within-trial neural indexing approach of EEG dynamics. We showed that neural indices of RL during acquisition were predictive of subsequent retention in the stimulus–response retention, even after controlling for set size. This result supports the key model prediction that RL processes during learning, which are stronger under high WM load, are responsible for increasing policy retention when WM is no longer available. In contrast, neural indices of RL were not predictive of performance in the reward retention test.

This result supports theoretical and empirical studies suggesting that model-free learning in the brain (especially the corticostriatal system) directly learns a stimulus–response policy using prediction errors from another system (actor-critic; Collins and Frank, 2014; Klein et al., 2017; Jaskir and Frank, 2023). By this account, the actor selecting policies would have no direct access to experienced reward values but only the propensity for a specific response for each of them. Participants could plausibly access their critic values for each stimulus and compare them in the reward retention phase, but they would not have had to do so during learning. Indeed, participants show above chance performance in such discriminations but only subtly (accuracy rises up to 60% at best); in contrast, accuracy in the stimulus–response retention test, which directly assesses what the actor would have learned, is far superior (~80% for the higher set sizes), despite being tested with further delays since learning.

For most simple RL tasks, these two classes of model-free RL algorithms (those that focus on learning expected values and the actor-critic), are largely indistinguishable as they both predict that an agent progressively chooses those actions that maximize reward. However, several theoretical and empirical studies suggest that the basic RL system in humans satisfies predictions of an actor-critic in behavior, imaging, and in theoretical models of corticostriatal contributions to RL (Li and Daw, 2011; Gold et al., 2012; Collins and Frank, 2014; Klein et al., 2017; Geana et al., 2022; Jaskir and Frank, 2023). Moreover, the model fits here did not improve if we allowed the Q learning agent to learn the difference between two versus one point and instead suggested that participants learned to simply maximize task performance, which effectively makes Q learning equivalent to an actor-critic at the level of task performance. Nevertheless, Q learners would, at minimum, learn the reward value of a stimulus in terms of the percentage of times they were correct (i.e., whether they got one or two points versus zero). Yet, the EEG marker of RL is still not related to performance in the reward retention test even when a correct performance there would be counted as simply choosing the stimulus that had yielded higher proportion of correct responses. Although our neural RL index cannot distinguish between an EEG metric of Q values or actor weights, the findings that it only predicts performance in the stimulus–response test provides initial evidence supporting the actor interpretation where the neural RL index reflects the policy rather than its reward value.

Although we focused mainly on how the RLWM mechanism informs retention, we also tested whether the interaction between RL and WM can be modulated by acute stress. Stress is known to have a major impact on learning and decision-making processes (Starcke and Brand, 2012; Raio et al., 2017; Cremer et al., 2021). Previous work had shown that acute stress alters prefrontal cortex functioning, thus impairing executive control over cognition (cognitive inhibition, task switching, working memory maintenance; Schwabe et al., 2011; Schwabe and Wolf, 2011; Plessow et al., 2012; Hamilton and Brigman, 2015; Bogdanov and Schwabe, 2016; Vogel et al., 2016; Goldfarb et al., 2017; Brown et al., 2020). On the other hand, acute stress was also shown to increase striatal dopamine activity (Vaessen et al., 2015) leading to better working-memory updating (Goldfarb et al., 2017) and improving executive control over motor actions (i.e., response inhibition; Schwabe and Wolf, 2012; Leong and Packard, 2014). We, therefore, predicted that stress would affect the WM versus RL trade-off such that it will impede the contribution of WM to learning and will instead enhance the relative contribution of RL computations. Current results did not confirm this hypothesis as only subtle differences were observed between the stress and control groups during the learning task and at the tests.

It is possible that the 25 min delay between the stressor and the beginning of the learning task hindered the stress response on behavior as it was previously suggested that both noradrenaline and cortisol levels need to be elevated in order for stress to affect WM performance (Elzinga and Roelofs, 2005; Roozendaal, et al., 2006; Barsegyan et al., 2010). Another intriguing possibility is that individuals with higher WM capacity were more resilient against cognitive impairments induced by stress and were also less biased toward habitual decision-making (Otto et al., 2013; Quaedflieg et al., 2019; Cremer et al., 2021). Future work should test directly the specific effect of stress on WM and RL interactions while taking into account participants' WM capacity as a factor.

To conclude, our results contribute to a better understanding of the coupled mechanism of WM and RL that can dynamically shift between relying more on the effortful, but fast and reliable WM system or the slow, more error-prone RL system that has retention benefits. We reported trial-by-trial evidence in the neural signal for this trade-off during learning and showed that greater reliance on the RL system when WM is degraded (i.e., when WM load is high) predicted better memory retention of learned stimulus–response associations. An intriguing possibility that remains to be tested is that the shift between the two systems is strategic and can be modulated by one's preference or ability to maximize immediate learning versus retention. However, it remains to be seen whether clinical populations with impairments in one or both systems of WM and RL might alter the flexible shifting between the two systems, possibly biasing the use of one system more than the other even when it is less advantageous.

References

Arnsten AF (2009) Stress signalling pathways that impair prefrontal cortex structure and function. Nat Rev Neurosci 10:410–422.

Barsegyan A, Mackenzie SM, Kurose BD, McGaugh JL, Roozendaal B (2010) Glucocorticoids in the prefrontal cortex enhance memory consolidation and impair working memory by a common neural mechanism. Proc Natl Acad Sci U S A 107:16655–16660.

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. J Stat Softw 67:1–48.

Bogdanov M, Schwabe L (2016) Transcranial stimulation of the dorsolateral prefrontal cortex prevents stress-induced working memory deficits. J Neurosci 36:1429–1437.

Brown TI, Gagnon SA, Wagner AD (2020) Stress disrupts human hippocampal-prefrontal function during prospective spatial navigation and hinders flexible behavior. Curr Biol 30: 1821–1833.e8.

Collins AG (2018) The tortoise and the hare: interactions between reinforcement learning and working memory. J Cogn Neurosci 30:1422–1432.

Collins AG, Frank MJ (2012) How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. Eur J Neurosci 35:1024–1035.

Collins AG, Frank MJ (2014) Opponent actor learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. Psychological Review 121:337–366.

Collins AG, Frank MJ (2018) Within-and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. Proc Natl Acad Sci U S A 115:2502–2507.

Collins AG, Ciullo B, Frank MJ, Badre D (2017a) Working memory load strengthens reward prediction errors. J Neurosci 37:4332–4342.

Collins AG, Albrecht MA, Waltz JA, Gold JM, Frank MJ (2017b) Interactions among working memory, reinforcement learning, and effort in value-based choice: a new paradigm and selective deficits in schizophrenia. Biol Psychiatry 82:431–439.

Cremer A, Kalbe F, Gläscher J, Schwabe L (2021) Stress reduces both model-based and model-free neural computations during flexible learning. Neuroimage 229:117747.

Delorme A, Makeig S (2004) EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods 134:9–21.

Elzinga BM, Roelofs K (2005) Cortisol-induced impairments of working memory require acute sympathetic activation. Behav Neurosci 119:98–103.

Frank MJ, Samanta J, Moustafa AA, Sherman SJ (2007) Hold your horses: impulsivity, deep brain stimulation, and medication in parkinsonism. Science 318:1309–1312.

Hamilton DA, Brigman JL (2015) Behavioral flexibility in rats and mice: contributions of distinct frontocortical regions. Genes Brain Behav 14:4–21.

Geana A, Barch DM, Gold JM, Carter CS, MacDonald III AW, Ragland JD, Silverstein SM, Frank MJ (2022) Using computational modeling to capture schizophrenia-specific reinforcement learning differences and their implications on patient classification. Biol Psychiatry Cogn Neurosci Neuroimaging 7:1035–1046.

Gold JM, Waltz JA, Matveeva TM, Kasanova Z, Strauss GP, Herbener ES, Collins AGE, Frank MJ (2012) Negative symptoms and the failure to represent the expected reward value of actions: behavioral and computational modeling evidence. Arch Gen Psychiatry 69:129–138.

Goldfarb EV, Froböse MI, Cools R, Phelps EA (2017) Stress and cognitive flexibility: cortisol increases are associated with enhanced updating but impaired switching. J Cogn Neurosci 29:14–24.

Jaskir A, Frank MJ (2023) On the normative advantages of dopamine and striatal opponency for learning and choice. Elife 12:e85107.

Kim J, Lee H, Han J, Packard M (2001) Amygdala is critical for stress-induced modulation of hippocampal long-term potentiation and learning. J Neurosci 21:5222–5228.

Klein TA, Ullsperger M, Jocham G (2017) Learning relative values in the striatum induces violations of normative decision making. Nature Commun 8:16033.

Leong KC, Packard MG (2014) Exposure to predator odor influences the relative use of multiple memory systems: role of basolateral amygdala. Neurobiol Learn Mem 109:56–61.

Li J, Daw ND (2011) Signals in human striatum are appropriate for policy update rather than value prediction. J Neurosci 31:5504–5511.

Lopez-Calderon J, Luck SJ (2014) ERPLAB: an open-source toolbox for the analysis of event-related potentials. Front Hum Neurosci 8:213.

Matuschek H, Kliegl R, Vasishth S, Baayen H, Bates D (2017) Balancing Type I error and power in linear mixed models. J Mem Lang 94:305–315.

Meier JK, Staresina BP, Schwabe L (2022) Stress diminishes outcome but enhances response representations during instrumental learning. Elife 11: e67517.

Oberauer K, Farrell S, Jarrold C, Lewandowsky S (2016) What limits working memory capacity? Psychol Bull 142:758–799.

Otto AR, Raio CM, Chiang A, Phelps EA, Daw ND (2013) Working-memory capacity protects model-based learning from stress. Proc Natl Acad Sci U S A 110:20941–20946.

Palminteri S, Khamassi M, Joffily M, Coricelli G (2015) Contextual modulation of value signals in reward and punishment learning. Nature Commun 6:8096.

Plessow F, Kiesel A, Kirschbaum C (2012) The stressed prefrontal cortex and goal-directed behaviour: acute psychosocial stress impairs the flexible implementation of task goals. Exp Brain Res 216:397–408.

Quaedflieg CWEM, Stoffregen H, Sebalo I, Smeets T (2019) Stress-induced impairment in goal-directed instrumental behaviour is moderated by baseline working memory. Neurobiol Learn Mem 158:42–49.

Rac-Lubashevsky R, Frank MJ (2021) Analogous computations in working memory input, output and motor gating: electrophysiological and computational modeling evidence. PLoS Comput Biol 17:e1008971.

Raio CM, Hartley CA, Orederu TA, Li J, Phelps EA (2017) Stress attenuates the flexible updating of aversive value. Proc Natl Acad Sci U S A 114:11241–11246.

Roozendaal B, Okuda S, De Quervain DF, McGaugh JL (2006) Glucocorticoids interact with emotion-induced noradrenergic activation in influencing different memory functions. Neuroscience 138:901–910.

Schwabe L, Schächinger H (2018) Ten years of research with the Socially Evaluated Cold Pressor Test: data from the past and guidelines for the future. Psychoneuroendocrinology 92:155–161.

Schwabe L, Wolf OT (2009) Stress prompts habit behavior in humans. J Neurosci 29:7191–7198.

Schwabe L, Wolf OT (2011) Stress-induced modulation of instrumental behavior: from goal-directed to habitual control of action. Behav Brain Res 219:321–328.

Schwabe L, Wolf OT (2012) Stress modulates the engagement of multiple memory systems in classification learning. J Neurosci 32:11042–11049.

Schwabe L, Haddad L, Schachinger H (2008) HPA axis activation by a socially evaluated cold-pressor test. Psychoneuroendocrinology 33:890–895.

Schwabe L, Höffken O, Tegenthoff M, Wolf OT (2011) Preventing the stress-induced shift from goal-directed to habit action with a β-adrenergic antagonist. J Neurosci 31:17317–17325.

Starcke K, Brand M (2012) Decision making under stress: a selective review. Neurosci Biobehav Rev 36:1228–1248.

Steyer R, Schwenkmezger P, Notz P, Eid M (1994) Testtheoretische Analysen der Mehrdimensionalen Befindlichkeitsfragebogens (MDBF). Diagnostica 40:320–328.

Vaessen T, Hernaus D, Myin-Germeys I, van Amelsvoort T (2015) The dopaminergic response to acute stress in health and psychopathology: a systematic review. Neurosci Biobehav Rev 56:241–251.

Vogel S, Fernández G, Joëls M, Schwabe L (2016) Cognitive adaptation under stress: a case for the mineralocorticoid receptor. Trends Cogn Sci 20:192–203.

Wimmer GE, Poldrack RA (2022) Reward learning and working memory: effects of massed versus spaced training and post-learning delay period. Mem Cognit 50:312–324.

Wirz L, Bogdanov M, Schwabe L (2018) Habits under stress: mechanistic insights across different types of learning. Curr Opin Behav Sci 20:9–16.

Yoo AH, Collins AG (2022) How working memory and reinforcement learning are intertwined: A cognitive, neural, and computational perspective. Journal of Cognitive Neuroscience 34:551–568.

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR PSYCHOLOGIE UND
BEWEGUNGSWISSENSCHAFT
Institut für Bewegungswissenschaft
Institut für Psychologie

Eidesstattliche Erklärung nach *(bitte Zutreffendes ankreuzen)*

□ **§ 7 (4) der Promotionsordnung des Instituts für Bewegungswissenschaft der Universität Hamburg vom 18.08.2010**

☒ **§ 9 (1c und 1d) der Promotionsordnung des Instituts für Psychologie der Universität Hamburg vom 20.08.2003**

Hiermit erkläre ich an Eides statt,

1. dass die von mir vorgelegte Dissertation nicht Gegenstand eines anderen Prüfungsverfahrens gewesen oder in einem solchen Verfahren als ungenügend beurteilt worden ist.
2. dass ich die von mir vorgelegte Dissertation selbst verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und keine kommerzielle Promotionsberatung in Anspruch genommen habe. Die wörtlich oder inhaltlich übernommenen Stellen habe ich als solche kenntlich gemacht.

Hamburg, 10.07.2024

Ort, Datum

Unterschrift

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR PSYCHOLOGIE UND
BEWEGUNGSWISSENSCHAFT
Institut für Bewegungswissenschaft
Institut für Psychologie

Erklärung gemäß ***(bitte Zutreffendes ankreuzen)***

□ **§ 4 (1c) der Promotionsordnung des Instituts für Bewegungswissenschaft der Universität Hamburg vom 18.08.2010**

☒ **§ 5 (4d) der Promotionsordnung des Instituts für Psychologie der Universität Hamburg vom 20.08.2003**

Hiermit erkläre ich,

ANNA CREMER (Vorname, Nachname),

dass ich mich an einer anderen Universität oder Fakultät noch keiner Doktorprüfung unterzogen oder mich um Zulassung zu einer Doktorprüfung bemüht habe.

Hamburg, 10.07.2024
Ort, Datum

A. Cremer
Unterschrift