EPFL

Thèse n° 10 510

Unveiling the complexity of learning and decision-making

Présentée le 19 juin 2024

Faculté des sciences de la vie
Laboratoire de psychophysique
Programme doctoral en neurosciences

pour l'obtention du grade de Docteur ès Sciences

par

Wei-Hsiang LIN

Acceptée sur proposition du jury

Prof. D. N. A. Van De Ville, président du jury
Prof. M. Herzog, directeur de thèse
Prof. Ph. Tobler, rapporteur
Prof. Ph. Sterzer, rapporteur
Prof. A. Mathis, rapporteur

École polytechnique fédérale de Lausanne

2024

Acknowledgments

This thesis would not have been possible without the support and guidance of many individuals. First and foremost, I wish to express my deepest gratitude to my thesis supervisor, Prof. Michael Herzog, for his unwavering support and invaluable insights, extending beyond academic matters. I am also immensely thankful to Prof. Carmen Sandi for her critical perspectives on the social dominance project, and to Prof. Wulfram Gerstner for his enriching discussions and contributions to the reinforcement learning project.

Additionally, I am grateful to everyone at the LPSY lab for their stimulating ideas and the enjoyable atmosphere. Special thanks to Dario and Simona for their help and fruitful discussions. A particular acknowledgment goes to Janir, who has been a pillar of support from my very first day in the lab, offering guidance and support across all projects. I also extend my gratitude to Maya and Marina from Georgia, whose direct assistance was invaluable to my projects. I must thank Joao for his contributions to the social dominance project and Alireza for his dedication to the curiosity and reinforcement learning studies.

Lastly, I would like to take this opportunity to express my profound gratitude to my family: my parents, grandmother, and my brother, for their unwavering support throughout my studies. A special acknowledgment goes to my wife, Shu-Ching Lee, who has been my steadfast companion in this journey abroad. Her dedication, from offering valuable suggestions to managing hardships together and providing care, has been immeasurable. I am deeply appreciative of her support and love.

Lausanne, 6th February 2024

Abstract

Reinforcement learning (RL) is crucial for learning to adapt to new environments. In RL, the prediction error is an important component that compares the expected and actual rewards. Dopamine plays a critical role in encoding these prediction errors. In my thesis, I investigate how human behavior in RL can be explained by RL models that incorporate factors such as novelty and abnormalities in the dopamine system. To broaden our understanding of RL and decision-making, I extend the study to include neural correlates of social dominance, examining their impact on decision-making processes. First, our model suggests that humans tend to adopt novelty-seeking strategies in RL tasks. Indeed, behaviorally, participants are often distracted by the emergence of novel states, especially among those who are more optimistic about potential rewards. Given the critical role of dopamine in RL, we examined its impact on two populations typically associated with dopamine dysregulation: patients with schizophrenia and the elderly. Our model suggests that restricted dopamine levels lead to deficits in reward-based learning due to poor encoding of the reward prediction error. On the other hand, the model predicts that punishment-based learning abilities should be preserved or even enhanced compared to controls. We could verify this prediction. Older adults demonstrate robust RL capabilities, with no significant differences compared to young adults, challenging the common belief about cognitive decline in healthy ageing. Finally, we investigate the neural correlates of social dominance in female participants. Our findings reveal an EEG component in dominant females similar to the previously reported one in dominant males, suggesting a potential universal neuromarker for social dominance during decision-making scenarios.

Keywords: *Reinforcement learning, decision-making, aging, schizophrenia, dopamine, social dominance, EEG, neuromarker*

Résumé

Le reinforcemnet learning (RL) est essentiel pour apprendre à s'adapter à de nouveaux environnements. Dans le RL, l'erreur de prédiction est un composant important qui compare les récompenses attendues et réelles. La dopamine joue un rôle crucial dans le codage de ces erreurs de prédiction. Dans ma thèse, j'examine comment le comportement humain dans le RL peut être expliqué par des modèles de RL qui intègrent des facteurs tels que la nouveauté et les anomalies du système dopaminergique. Pour élargir notre compréhension du RL et de la prise de décision, j'étends l'étude pour inclure les corrélats neuronaux de la dominance sociale, en examinant leur impact sur les processus de prise de décision. Tout d'abord, notre modèle suggère que les humains ont tendance à adopter des stratégies de recherche de nouveauté dans les tâches de RL. En effet, d'un point de vue comportemental, les participants sont souvent distraits par l'émergence de nouveaux états, surtout parmi ceux qui sont plus optimistes quant aux récompenses potentielles. Étant donné le rôle crucial de la dopamine dans le RL, nous avons examiné son impact sur deux populations généralement associées à une dysrégulation de la dopamine : les patients atteints de schizophrénie et les personnes âgées. Notre modèle suggère que des niveaux restreints de dopamine entraînent des déficits dans l'apprentissage basé sur la récompense en raison d'un mauvais codage de l'erreur de prédiction de la récompense. D'autre part, le modèle prédit que les capacités d'apprentissage basées sur la punition devraient être préservées ou même améliorées par rapport aux contrôles. Nous avons pu vérifier cette prédiction. Les adultes plus âgés démontrent des capacités robustes de RL, sans différences significatives par rapport aux jeunes adultes, remettant en question la croyance commune sur le déclin cognitif dans le vieillissement sain. Enfin, nous étudions les corrélats neuronaux de la dominance sociale chez les participantes féminines. Nos résultats révèlent un composant EEG chez les femmes dominantes similaire à celui précédemment rapporté chez les hommes dominants, suggérant un neuromarqueur universel potentiel pour la dominance sociale lors de scénarios de prise de décision.

Mots-clés: *Apprentissage par renforcement, prise de décision, vieillissement, schizophrénie, dopamine, dominance sociale, EEG, neuromarqueur*

Contents

1. Preface

My Ph.D. research aimed to investigate various levels of RL and decision-making by focusing on diverse populations (patients, ageing, females) and utilizing diverse methodologies (behavior, EEG). In study I, we examined the impact of stochastic states on RL process among healthy participants and proposed an algorithm to explain human behavior. Conversely, in study II, we examined RL in patients with schizophrenia to illustrate the influence of abnormalities in the dopamine system on the RL process. As a bridging exploration, in study III, we assessed RL in an ageing population to understand potential differences between older and young adults, especially considering that the dopamine system in older individuals is not deficient but rather degraded. Finally, in study IV, we turned our attention to the intrinsic personality trait of social dominance. Our objective was to identify a universal neural marker for social dominance, with a specific emphasis on the female population, as it has not been studied in this context before.

Study I: The curse of optimism: a persistent distraction by novelty.

(1) Modirshanechi, A., **Lin, W.-H.**, Xu, H.A., Herzog, M.H., and Gerstner, W. The curse of optimism: a persistent distraction by novelty (manuscript submitted)

In this study, we aimed to understand how human learning is affected by novelty and internal beliefs about rewards. Beyond behavioral testing, we utilized computational modeling to determine which RL model best explains human behavior.

My role in this study included implementing the experiment and consolidating all experiment-related information. I also contributed to the discussion of the manuscript.

Study II: Abnormal dopamine levels lead to deteriorated reward but enhanced punishment learning in schizophrenia patients

(2) **Lin, W.-H.**, Roinishvili, M., Okruashvili, M., Chkonia, E., and Herzog, M.H. Abnormal dopamine levels lead to deteriorated reward but enhanced punishment learning in schizophrenia patients. (manuscript submitted)

Here, we proposed an explanation, using the Q-learning framework, for how an abnormal dopamine system affects RL performance in schizophrenia patients in the domain of reward and punishment. We validated our theory through both behavioral experiments and simulation.

I was responsible for data analysis and manuscript drafting for this project.

Study III: Intact reinforcement learning in healthy ageing

(3) **Lin W.-H.**, Pilz, K.S., Herzog, M.H., and Kunchulia, M. Intact reinforcement learning in healthy ageing (manuscript submitted).

In this study, we conducted RL experiments to explore differences in the RL abilities of older and younger adults, as existing literature does not offer uniformly conclusive findings on this topic.

I was responsible for data analysis and manuscript drafting for this project.

Study IV: Behavioral and Neural Markers of Social Dominance: A Female-Focused Perspective

(4) **Lin, W.-H.**, da Cruz, J.R., Sandi, C., and Herzog M.H., and. Behavioral and neural markers of social dominance: A female-focused perspective. (in preparation)

In this study, we investigated whether dominant females respond more quickly in decision-making tasks. Additionally, we explored the presence of any EEG features that can be considered as neuromarkers for social dominance.

I collected and analyzed the data and was responsible for drafting the manuscript for this project.

Other publications not included in this thesis:

(5) Gordillo, D., da Cruz, J.R., Chkonia, E., **Lin, W.-H.**, Favrod, O., Brand, A., Figueiredo, P., Roinishvili, M., and Herzog, M.H. (2022). The EEG multiverse of schizophrenia. Cereb. Cortex, bhac309.

I conducted the frequency, connectivity, and network analyses on the resting-state EEG data.

(6) Garobbio, S., **Lin, W.-H.**, Kunchulia, M., Herzog, M.H. High correlations between neural correlates of visual tests, but no prediction of performance. (in preparation)

I conducted the EEG analysis, focusing on extracting various ERP components and variance features from the EEG data, and also took responsibility for drafting the manuscript.

2. Introduction

2.1 Reinforcement learning

RL is mainly concerned with identifying the optimal policy for a given situation, which is often similar to navigating in an environment. Consider being a pioneer in an unfamiliar territory, navigating through unknown places in order to search for hidden resources or treasures. Initially, you can merely choose the direction randomly. However, upon unexpectedly discovering a reward, you not only update this information to memory, but also start to form patterns or strategies for further exploration. This is just an example for the classic situation (Sutton & Barto, 2018): an agent changes from one location in an environment (state) to another one by choosing actions to find one or more goal states.

Essentially, goal information is updated through the "prediction error", which is the difference between the anticipated and actual reward. For example, finding unexpectedly a treasure in a certain location causes a significant prediction error since there was no expectation to find the treasure. The prediction error serves as the learning signal in RL. A classic algorithm that integrates all these elements is the Q-learning model (Watkins & Dayan, 1992). This model updates the value of each state-action pair based on the strength of the prediction error (formula 1).

$$Q(S,A) \leftarrow Q(S,A) + \alpha \left(r + \max_i Q(S_{new}, A_i) - Q(S,A) \right) \quad \text{formula 1}$$

Numerous studies in humans, primates, and rodents tried to investigate the neural mechanism of the prediction error (Gläscher et al., 2010; Schultz et al., 1997). Evidence suggests that dopaminergic neurons play a pivotal role in encoding prediction error, with these neurons exhibiting greater activity in response to unexpected rewards compared to expected ones (Fig. 2.1). Subsequently, we will explore RL across various populations and investigate factors that may influence individual behavior.

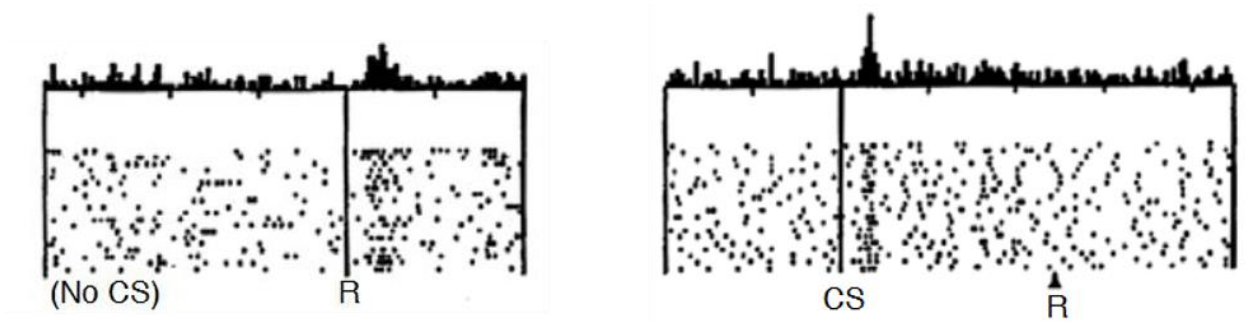

Figure 2.1. Neural evidence of prediction error. Schultz et al. (1997) examined neural activity of dopamine neurons in primates. When an unexpected reward was presented (R in the figure), an increase in dopaminergic neuron activity was observed (left figure). Interestingly, when the acquisition of a reward was predicted, no such increase in activity was observed (right figure). This implies that the observed activity is not encoding the reward itself but rather the reward prediction error.

2.2 Curiosity and novelty

In addition to fundamental elements of RL, such as prediction error, other factors can influence learning. These include intrinsic motivations like curiosity (Gruber et al., 2019). Intrinsic motivation to engage in an environment is a crucial factor. Various studies have shown that curiosity triggers species to explore their surroundings (Berlyne, 1950). This concept can also be applied to the field of machine learning. Curiosity-driven RL algorithms are also considered essential to facilitate the learning of artificial intelligence (Burda et al., 2018). However, RL algorithms often don't account for scenarios where agents get stuck in non-rewarding states, potentially due to factors like noise or novelty (Aubret et al., 2019). An agent based on an algorithm, such as information-gain, can efficiently escape from this situation since, over time, there is no more information to acquire. Conversely, algorithms such as novelty-seeking might struggle since they can be continuously attracted by new stimuli, even if these are not rewarding (Aubret et al., 2019). Therefore, it would be valuable to further understand how humans behave in these contexts and to determine whether any existing algorithm align closely with human responses.

2.3 Reinforcement learning and schizophrenia

Dopamine is, as previously mentioned, a crucial neurotransmitter in RL and plays a vital role in encoding prediction error. Consequently, abnormalities in the dopamine system can significantly impact RL abilities. These abnormalities are critical in several psychiatric diseases,

especially in schizophrenia. For instance, dopamine D2 antagonists can reduce psychotic symptoms in schizophrenia patients, while dopamine agonists can induce the symptoms, even in healthy individuals. A key factor is the chronically elevated baseline dopamine levels observed in such patients, suggesting a relationship between the deficits in dopamine levels and psychiatric symptoms.

Beyond these physiological aspects, patients with schizophrenia often demonstrate a range of cognitive deficits (Kuperberg & Heckers, 2000). Furthermore, they exhibit different types of symptoms, both positive and negative. Some studies have demonstrated that schizophrenia patients have deficits in RL. Specifically, these patients perform suboptimally in many tasks and exhibit greater perseveration, a persistent repetition of specific behaviors, a pattern commonly observed in these patients (A. G. Collins et al., 2014; Strauss et al., 2011; Waltz et al., 2007a). Such findings highlight the significance of studying schizophrenia to gain insights into how dopamine affects RL.

2.4 Reinforcement learning and ageing

Aging is a primary focus for many researchers, especially when investigating its impact on cognitive functions. While various studies suggest that learning abilities, including RL, deteriorates with age (Anguera et al., 2011; Salthouse, 2009; van de Vijver & Ligneul 2020), findings are not always consistent. For instance, Daniel et al. (2020) observed that older adults performed well in simple RL tasks. However, when the task involved multi-dimensional stimuli to determine a reward, older adults exhibited learning deficits, possibly due to increased attentional demands in the task. In contrast, Lighthall et al. (2018) manipulated the inter-stimuli interval (ISI) in an RL task to modulate the level of memory load, and found no performance difference between older and young adults. The results suggest that RL abilities remain intact in older individuals. This implies that, in some contexts, RL capabilities may be unaffected in the elderly. Therefore, further studies are essential to reconcile these differing results.

2.5 Social dominance

Studies have shown that learning and decision-making are influenced not only by one's intrinsic abilities but also by external factors, such as the presence of others or the social context (Klucharev et al., 2009; Suzuki et al., 2012). Given this, it becomes intriguing to discern which

specific elements within a social context impact individual behavior. Among all the aspects, social dominance stands out as a potential key influencer. Though the exact definition of social dominance is elusive, it is often defined as one's relative power or rank within a social group (Rowell, 1974). In fact, social dominance is frequently associated with higher-level behaviors such as decision-making or leadership (Zink et al., 2008; Guinote 2017; Johnson et al. 2012).

However, an alternative perspective posits social dominance as a personality trait, suggesting that it might exist independently of any social context (Buss & Craik 1980). This raises the question: Is "social dominance" a characteristic that necessitates a social context for examination? da Cruz et al. (2018) used a questionnaire, rather than a social context to determine an individual's dominance level. Participants were then asked to complete several different decision-making tasks. The study demonstrated that dominant males made decisions more quickly. Moreover, using EEG recordings, they also identified an N2/P2 EEG component that exhibited increased activity in dominant males, suggesting a potentially neural marker for social dominance.

3. Study I: The curse of optimism: a persistent distraction by novelty

Curiosity is an essential factor in guiding the behavior of RL algorithms, and there is a substantial interest in understanding how it drives human learning. Some studies have proposed that curiosity works to reduce the uncertainty or stochasticity in the environment. For instance, research suggested that when children are presented with a novel toy without instructions, they exhibit increased engagement compared to a toy with instructions (Bonawitz et al., 2011; Kidd & Hayden, 2015). However, the pursuit of novelty may not always be optimal. Given that novelty can sometimes represent a reward-independent stochasticity, remaining in such a situation could preclude the opportunity to update rewards in alternative options (Aubret et al., 2019).

Nevertheless, no studies have directly investigated the factors and models that explain how the existence of such stochasticity might affect human behavior. Therefore, in the first study, we sought to address two key aspects: 1) whether humans are distracted by reward-independent stochastic stimuli, and 2) the factors driving this behavior. To test the first question, we designed a RL task in which participants see different images (state), each with three options below it. They have to select one of the options (action) in order to proceed to the next image (Fig. 3.1A). They were told that a hidden reward (goal) might appear in the environment, and they had to find the reward five times in total. There are two important manipulations in the study. First, each of the participants were informed that three different levels of reward (low, medium, and high) can be found in the environment, thereby manipulating participants' belief about potential payouts. Additionally, within the environment, we introduced a subregion continuously presenting novel states to the participants. This subregion acted as a "stochastic", offering no rewards (Fig. 3.1B).

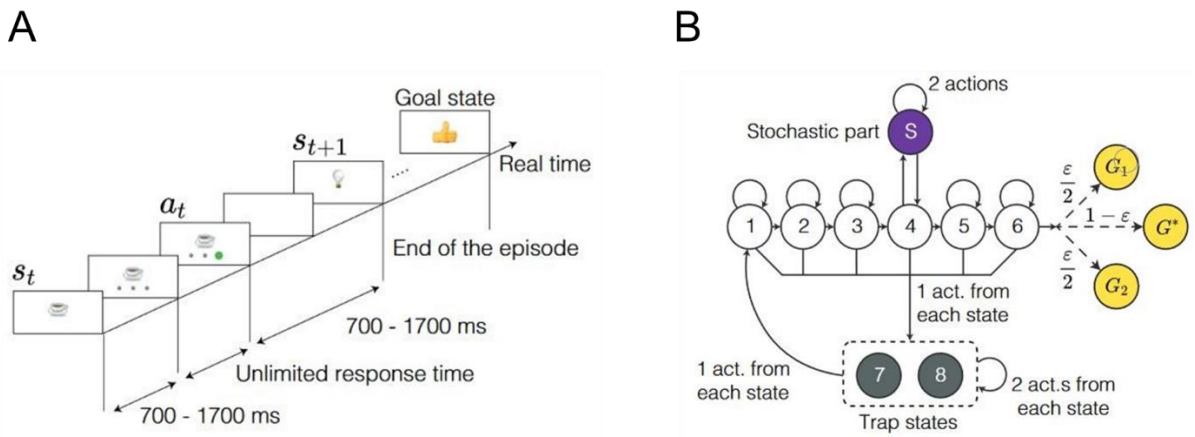

Figure 3.1. Experimental design of the RL task. (A) Elements of an episode. Participants were presented with an image accompanied by three gray disks below it. They had to press one of the buttons to transition to another image, continuing until they found the goal state, representing the end of an episode. (B) Structure of the RL environment. The environment incorporates a stochastic component where two of the actions consistently lead to the presentation of a new image, creating stochastic conditions for the participants.

Our findings revealed that participants exhibited suboptimal performance, choosing to remain in novelty states. Notably, this behavior seemed to be influenced by the size of the anticipated reward, the smaller the expected monetary reward, the longer the participants chose to stay in novelty states. This addressed our first question, confirming that humans do exhibit suboptimal behavior by persisting in stochastic yet novel states. Furthermore, this suboptimal behavior correlates with participants' beliefs about the payoffs available in the environment. Participants who were optimistic, believing in the possibility of high payoffs, were more prone to remain in the "stochastic" states, whereas those with less optimistic beliefs were more likely to avoid such conditions (Fig. 3.2A). Subsequently, we modeled human behavior to discern which RL models could most accurately capture these observed behaviors. We explored three possible models, novelty-seeking, surprise-seeking, and information-gain-seeking. Our results indicated that the novelty-seeking model provided the best fit for the observed human behaviors compared to the other models (Fig. 3.2B). This implies that the pursuit of novelty might be a key intrinsic motivation causing individuals to remain in suboptimal states.

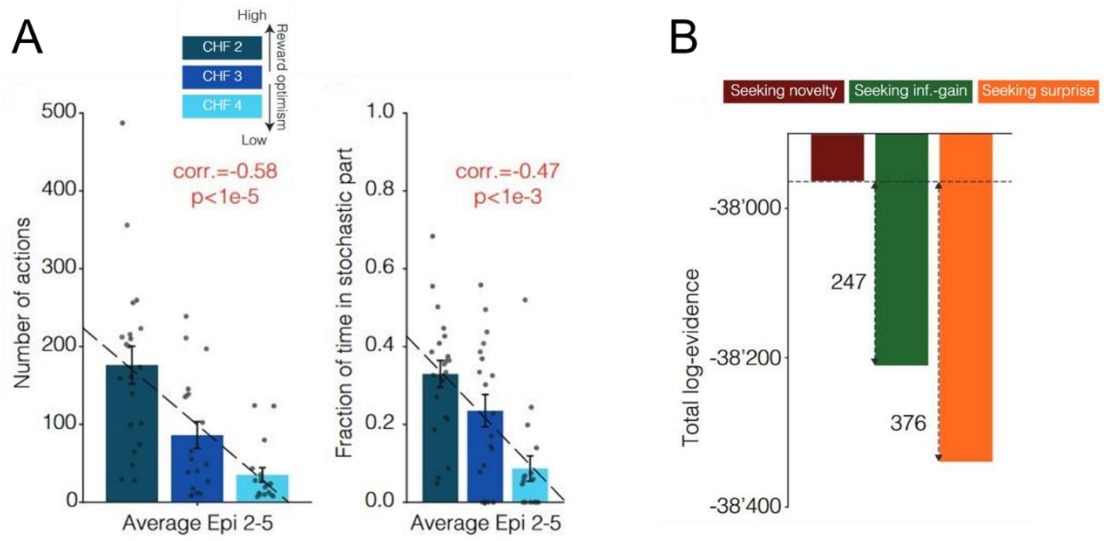

Figure 3.2. Results of behavioral analysis and model fitting. (A) Left: The number of actions performed in the RL task, with the three bars indicating the different monetary rewards received by the participants. Right: The fraction of time spent in stochastic part, with the y-axis showing the proportion of actions taken within the stochastic states. (B) The log-likelihood of the model fitting. We computed the log-likelihood to

determine which model best explained the observed behavior, with the bars representing three different models.

The curse of optimism: a persistent distraction by novelty

Alireza Modirshanechi[1,2,*], Wei-Hsiang Lin[1], He A. Xu[1], Michael H. Herzog[1], and Wulfram Gerstner[1,2]

1. Brain-Mind Institute, School of Life Sciences, EPFL, Lausanne, Switzerland
2. School of Computer and Communication Sciences, EPFL, Lausanne, Switzerland

∗ Corresponding author: alireza.modirshanechi@epfl.ch

3.1 *Abstract*

Human curiosity has been interpreted as a drive for exploration and modeled by intrinsically motivated reinforcement learning algorithms. An unresolved challenge in machine learning is that several of these algorithms get distracted by reward-independent stochastic stimuli. Here, we ask whether humans get distracted by the same stimuli as the algorithms. We design an experimental paradigm where human participants search for rewarding states in an environment with a highly 'stochastic' but reward-free sub-region. We show that (i) participants get repeatedly and persistently distracted by novelty in the stochastic part of the environment; (ii) optimism about the availability of other rewards increases this distraction; and (iii) the observed distraction pattern is consistent with the predictions of algorithms driven by novelty but not with 'optimal' algorithms driven by information-gain. Our results suggest that humans use suboptimal but computationally cheap curiosity-driven policies for exploration in complex environments.

3.2 *Introduction*

Curiosity drives humans and animals to explore their environments[1–3] and to search for potentially more valuable sources of reward (e.g., more nutritious foods or better-paid jobs) than those currently available[4,5]. In computational neuroscience and psychology, intrinsically motivated reinforcement learning (RL) algorithms[6,7] have been proposed as models of curiosity-driven behavior[8–11] with novelty, surprise, or information-gain as intrinsic motivation in addition to the extrinsic motivation by nutritious or monetary reward[11]. These algorithms have been successful not only in explaining aspects of exploration in humans and animals[12–18] but also in solving complex machine learning tasks with sparse or even no (extrinsic) rewards such as in computer games[19–21] and high-dimensional control problems[22–25]. Despite their successes, these algorithms face a serious challenge: Intrinsically motivated agents are prone to distraction by reward-independent stochasticity (the so-called 'noisy TV' problem)[26,27], i.e., they are attracted to novel, surprising, or just noisy states independently of whether or not these states are rewarding[28].

The extent of distraction varies between different algorithms, and designing efficient noise-robust algorithms is an ongoing line of research in machine learning [29–32]. In particular, it is well-known that artificial RL agents seeking information-gain eventually lose their interest in stochasticity when exploration yields no further information, whereas RL agents seeking surprise or novelty exhibit a persistent attraction by stochasticity[26,27,33]. Here, we ask (i) whether humans get distracted in the same situations as intrinsically motivated RL agents and, if so, (ii) whether this distraction vanishes (similar to seeking information-gain) or persists (similar to seeking surprise or novelty) over time.

To answer these questions, we bring ideas from machine learning[26,27] to behavioral neuroscience and design a novel experimental paradigm with a highly stochastic part. We test the predictions of three different intrinsically motivated algorithms (i.e., driven by novelty, surprise, and information-gain) against the behavior of human participants and show that human behavior is both qualitatively and quantitatively consistent with that of novelty-driven RL agents: Human participants exhibit a persistent distraction by novelty, and the degree of this distraction correlates with their degree of 'reward optimism', where reward optimism is defined by the experimental procedure. Our results provide evidence for (i) novelty-driven RL algorithms as models of human curiosity even when novelty-seeking is suboptimal and (ii) the influence of reward optimism on the relative importance of novelty-seeking versus reward-seeking in human decision making.

3.3 *Methods*

Ethics Statement

The data for human experiment were collected under CE 164/2014, and the protocol was approved by the 'Commission cantonale d''ethique de la recherche sur l''ˆetre humain'. All participants were informed that they could quit the experiment at any time, and they all signed a written informed consent. All procedures complied with the Declaration of Helsinki (except for pre-registration).

Experimental procedure for human participants

63 participants joined the experiment. Data of 6 participants were removed (see below) and, thus, data of 57 participants (27 females, mean age 24.1 ± 4.1 years) were included in the analyses. All participants were naïve to the purpose of the experiment and had normal or corrected-to-normal visual acuity. The experiment was scripted in MATLAB using the Psychophysics Toolbox[96].

Before starting the experiment, the participants were informed that they need to find either one of the 3 goal states 5 times. They were shown the 3 goal images and informed that different images had different reward values of 2 CHF, 3 CHF, and 4 CHF. Specifically, they were given the example that 'if you find the 2 CHF goal twice, 3 CHF goal once, and 4 CHF goal twice, then you will be paid $2 \times 2 + 1 \times 3 + 2 \times 4 = 15$ CHF'; see 'Informing RL agents of different goal states and modeling optimism' for how simulated efficient agents and simulated participants were given this information. At each trial, participants were presented an image (state) and three grey disks below the image (Fig. 3.4.1C). Clicking on a disk (action) led participants to a subsequent image which was chosen based on the underlying graph of the environment in Fig. 3.4.1A-B (which was unknown to the participants). Participants clicked through the environment until they found one of the goal states which finished an episode (Fig. 3.4.1C).

The assignment of images to states and disks to actions was random but kept fixed throughout the experiment and among participants. Exceptionally, we did not make the assignment for the actions in state 4 before the start of the experiment. Rather, for each participant, we assigned the disk that was chosen in the 1st encounter of state 4 to the stochastic action and the other two disks randomly to the bad and progressing actions, respectively (Fig. 3.4.1A). With this assignment,

we made sure that all human participants would visit the stochastic part at least once during episode 1. The same protocol was used for simulated efficient agents and simulated participants.

Before the start of the experiment, we randomly assigned the different goal images (corresponding to the three reward values) to different goal states G^*, G1, and G2 in Fig. 3.4.1A, separately for each participant. The image and hence the reward value were then kept fixed throughout the experiment. In other words, we randomly assigned different participants to different environments with the same structure but different assignments of reward values. We, therefore, ended up with 3 groups of participants: 23 in the 2 CHF group, 20 in the 3 CHF group, and 20 in the 4 CHF group. The probability of encountering a goal state other than G^* is controlled by the parameters ε. We considered ε to be around machine precision 10^{-8}, so we have $(1-\varepsilon)^{5\times 63} \approx 1 - 10^{-5} \approx 1$, meaning that all 63 participants would be taken almost surely to the goal state G^* in all 5 episodes. We note, however, that a participant could in principle observe any of the 3 goals if they could choose the progressing action at state 6 sufficiently many times because $\lim_{t\to\infty}(1-\varepsilon)^t = 0$.

2 participants (in the 2 CHF group) did not finish the experiment, and 4 participants (1 in the 3 CHF group and 3 in the 4 CHF group) took more than 3 times group-average number of actions in episodes 2-5 to finish the experiment. We considered this as a sign of being non-attentive and removed these 6 participants from further analyses.

The sample size was determined by a power analysis performed on the data of the efficient simulations done for Fig. 3.4.2 (see 'Efficient model-based planning for simulated participants' for the simulation details). Our goal was to have a statistical power of more than 80% (with a significance level of 0.05) for correlations in panels Fig. 3.4.2A, C, and E as well as for the differences for the highest importance of intrinsic rewards in Fig. 3.4.2D and F.

The correction for multiple hypotheses testing was done by controlling the False Discovery Rate at 0.05[50] over all 22 null hypotheses that are tested in Fig. 3.4.3, Fig. 3.4.5, and Fig. 3.4.6 (p-value threshold: 0.034). All Bayes Factors (abbreviated BF in the figures) were evaluated using Schwartz approximation[49] to avoid any assumptions on the prior distribution. We note that evaluating the Bayes Factors using priors suggested by ref.[97,98] does not change our conclusions. We also note that using the Spearman correlation instead of the Pearson correlation in Fig. 3.4.2A, C, and E, Fig. 3.4.3A, and Fig. 3.4.6A does not change our conclusions.

Full hybrid model

We first present the most general case of our algorithm as visualized in Fig. 3.4.1D and then explain the special cases used for simulating efficient agents (Fig. 3.4.2) and for modeling human behavior (Fig. 3.4.4-Fig. 3.4.6). We used ideas from non-parametric Bayesian inference[99] to design an intrinsically motivated RL algorithm for environments where the total number of states is unknown. We present the final results here and present the derivations and pseudo-code in Supplementary Materials.

We indicate the sequence of actions and states until time t by $s_{1:t}$ and $a_{1:t}$, respectively, and define the set of all known states at time t as

$$\mathcal{S}^{(t)} = \{s: \exists t' \in \{1, \dots, t\} \text{ s.t. } s = s_{t'}\} \cup \{\tilde{G}_0, \tilde{G}_1, \tilde{G}_2\} \quad (1)$$

where $\tilde{G}_i$ s are our three different goal states $-\tilde{G}_0$ corresponds to the 2CHF goal, $\tilde{G}_1$ to the 3CHF goal, and $\tilde{G}_2$ to the 4CHF goal. Note that $\{\tilde{G}_0, \tilde{G}_1, \tilde{G}_2\}$ represents the images of the goal states and not their locations G^*, G_1, and G_2 and that the assignment of images to locations is unknown to the model. Hence, since $t = 0$, the simulated efficient agents and the simulated participants are aware of the existence of multiple goal states in the environment. In a more general setting, $\{\tilde{G}_0, \tilde{G}_1, \tilde{G}_2\}$ should be replaced by the set of all states whose images were shown to participants prior to the start of the experiment. After a transition to state $s_{t+1} = s'$ resulting from taking action $a_t = a$ at state $s_t = s$, the reward functions R_{ext} and $R_{\text{int},t}$ evaluate the reward values $r_{\text{ext},t+1}$ and $r_{\text{int},t+1}$. We define the extrinsic reward function R_{ext} as

$$R_{\text{ext}}(s, a \to s') = \delta_{s',\tilde{G}_0} + r_1^* \delta_{s',\tilde{G}_1} + r_2^* \delta_{s',\tilde{G}_2} \quad (2)$$

where δ is the Kronecker delta function, and we assume (without loss of generality) a subjective extrinsic reward value of 1 for $\tilde{G}_0$ (2CHF goal) and subjective extrinsic reward values of $r_1^* \geq 1$ and $r_2^* \geq 1$ for $\tilde{G}_1$ and $\tilde{G}_2$, respectively. The prior information of human participants about the difference in the monetary reward values of different goal states can be modeled in simulated participants by varying r_1^* and r_2^* (see 'Informing RL agents of different goal states and modeling optimism'). We discuss choices of $R_{\text{int},t}$ in the next section.

As a general choice for the RL algorithm in Fig. 1D, we consider a hybrid of model-based and model-free policy[18,36,38,52] The model-free (MF) component uses the sequence of states $s_{1:t}$, actions $a_{1:t}$, extrinsic rewards $r_{\text{ext }1:t}$, and intrinsic rewards $r_{\text{int},1:t}$ (in the two parallel branches in Fig. 1D) and estimates the extrinsic and intrinsic Q-values $Q^{(t)}_{\text{MF,ext}}$ and $Q^{(t)}_{\text{MF,int}}$, respectively.

Traditionally, MF algorithms do not need the total number of states[37], thus the MF component of our algorithm remains similar to that of previous studies[18,35] : At the beginning of episode 1 , we initialize Q-values at $Q^{(0)}_{\mathrm{MF,ext}}$ and $Q^{(0)}_{\mathrm{MF,\ int}}$. Then, the estimates are updated recursively after each new observation. After the transition $(s_t, a_t) \rightarrow s_{t+1}$, the agent computes extrinsic and intrinsic reward prediction errors $RPE_{\mathrm{ext},t+1}$ and $RPE_{\mathrm{int},t+1}$, respectively:

$$\begin{aligned} RPE_{\mathrm{ext},t+1} &= r_{\mathrm{ext},t+1} + \lambda_{\mathrm{ext}} V^{(t)}_{\mathrm{MF,ext}}(s_{t+1}) - Q^{(t)}_{\mathrm{MF,ext}}(s_t, a_t) \\ RPE_{\mathrm{int},t+1} &= r_{\mathrm{int},t+1} + \lambda_{\mathrm{int}} V^{(t)}_{\mathrm{MF,int}}(s_{t+1}) - Q^{(t)}_{\mathrm{MF,int}}(s_t, a_t), \end{aligned} \quad (3)$$

where λ_{ext} and $\lambda_{\mathrm{int}} \in [0,1)$ are the discount factors for extrinsic and intrinsic reward seeking, respectively, and $V^{(t)}_{\mathrm{MF,ext}}(s_{t+1}) = \max_{a'} Q^{(t)}_{\mathrm{MF,ext}}(s_{t+1}, a')$ and $V^{(t)}_{\mathrm{MF,\ int}}(s_{t+1}) = \max_{a'} Q^{(t)}_{\mathrm{MF,\ int}}(s_{t+1}, a')$ are the extrinsic and intrinsic V-values[37] of the state s_{t+1}, respectively. We use two separate eligibility traces[35,37] for the update of Q-values, one for extrinsic reward $e_{\mathrm{ext},t}$ and one for intrinsic reward $e_{\mathrm{int},t}$, both initialized at zero at the beginning of each episode. The update rules for the eligibility traces after taking action a_t at state s_t is

$$\begin{aligned} e_{\mathrm{ext},t+1}(s,a) &= \begin{cases} 1 & \text{if } s = s_t, a = a_t \\ \lambda_{\mathrm{ext}} \mu_{\mathrm{ext}} e_{\mathrm{ext},t}(s,a) & \text{otherwise} \end{cases} \\ e_{\mathrm{int},t+1}(s,a) &= \begin{cases} 1 & \text{if } s = s_t, a = a_t \\ \lambda_{\mathrm{int}} \mu_{\mathrm{int}} e_{\mathrm{int},t}(s,a) & \text{otherwise} \end{cases} \end{aligned} \quad (4)$$

where λ_{ext} and λ_{int} are the discount factors defined above, and μ_{ext} and $\mu_{\mathrm{int}} \in [0,1]$ are the decay factors of the eligibility traces for the extrinsic and intrinsic rewards, respectively. The update rule is then $\Delta Q^{(t+1)}_{\mathrm{MF}}(s,a) = \rho e_{t+1}(s,a) RPE_{t+1}$, where e_{t+1} is the eligibility trace (i.e., either $e_{\mathrm{ext},t+1}$ or $e_{\mathrm{int},t+1}$), RPE_{t+1} is the reward prediction error (i.e., either $RPE_{\mathrm{ext},t+1}$ or $RPE_{\mathrm{int},t+1}$), and $\rho \in [0,1)$ is the learning rate.

The model-based (MB) component builds a world-model that summarizes the structure of the environment by estimating the probability $p^{(t)}(s' \mid s,a)$ of the transition $(s,a) \rightarrow s'$. To do so, an agent counts the transition $(s,a) \rightarrow s'$ recursively and using a leaky integration[100,101]:

$$\tilde{C}^{(t+1)}_{s,a,s'} = \begin{cases} \kappa \tilde{C}^{(t)}_{s,a,s'} + \delta_{s',s_{t+1}} & \text{if } s = s_t, a = a_t \\ \tilde{C}^{(t)}_{s,a,s'} & \text{otherwise} \end{cases} \quad (5)$$

where δ is the Kronecker delta function, $\tilde{C}^{(0)}_{s,a,s'} = 0$, and $\kappa \in [0,1]$ is the leak parameter and accounts for imperfect memory and model-building in humans. If $\kappa = 1$, then $\tilde{C}^{(t+1)}_{s,a,s'}$ is the exact count of transition $(s,a) \to s'$. These counts are used to estimate the transition probabilities

$$p^{(t)}(s' \mid s,a) = \begin{cases} \frac{\epsilon_{\text{obs}} + \tilde{C}^{(t)}_{s,a,s'}}{\epsilon_{\text{new}} + \epsilon_{\text{obs}}|\mathcal{S}^{(t)}| + \tilde{C}^{(t)}_{s,a}} & \text{if } s' \in \mathcal{S}^{(t)} \\ \frac{\epsilon_{\text{new}}}{\epsilon_{\text{new}} + \epsilon_{\text{obs}}|\mathcal{S}^{(t)}| + \tilde{C}^{(t)}_{s,a}} & \text{if } s' = s_{\text{new}} \end{cases} \quad (6)$$

where $\tilde{C}^{(t)}_{s,a} = \sum_{s'} \tilde{C}^{(t)}_{s,a,s'}$ is the counts of taking action a at state s, $\epsilon_{\text{obs}} \in \mathbb{R}^+$is a free parameter for the prior probability of transition to a known state (i.e., states in $\mathcal{S}^{(t)}$), and $\epsilon_{\text{new}} \in \mathbb{R}^+$is a free parameter for the prior probability of transition to a new state (i.e., states not in $\mathcal{S}^{(t)}$) - see Supplementary Materials for derivations. Choosing $\epsilon_{\text{new}} = 0$ is equivalent to assuming there is no unknown state in the environment, for which the estimate in Eq. 6 is reduced to the classic Bayesian estimate of transition probabilities in bounded discrete environments[18,36]. The transition probabilities are then used in a novel variant of prioritized sweeping[37,40] adapted to deal with an unknown number of states. The prioritized sweeping algorithm computes a pair of Q-values, i.e., $Q^{(t)}_{\text{MB, ext}}$ for extrinsic and $Q^{(t)}_{\text{MB,int}}$ for intrinsic rewards, by solving the corresponding Bellman equations[37] with $T_{PS,\text{ext}}$ and $T_{PS,\text{int}}$ iterations, respectively for $Q^{(t)}_{\text{MB,ext}}$ and $Q^{(t)}_{\text{MB, int}}$. See Supplementary Material for details.

Finally, actions are chosen by a hybrid softmax policy[37] : The probability of taking action a in state s at time t is

$$\pi_t(a \mid s) \propto \exp[\beta_{\text{MB,ext}} Q^{(t)}_{\text{MB,ext}}(s,a) + \beta_{\text{MF,ext}} Q^{(t)}_{\text{MF,ext}}(s,a) + \beta_{\text{MB,int}} Q^{(t)}_{\text{MB,int}}(s,a) + \beta_{\text{MF,int}} Q^{(t)}_{\text{MF,int}}(s,a)], \quad (7)$$

where $\beta_{\text{MB,ext}} \in \mathbb{R}^+, \beta_{\text{MF, ext}} \in \mathbb{R}^+, \beta_{\text{MB, int}} \in \mathbb{R}^+$, and $\beta_{\text{MF, int}} \in \mathbb{R}^+$ are free parameters (i.e., inverse temperatures of the softmax policy[37]) expressing the contribution of each Q-value to actionselection. For Fig. 3.4.1D, we defined

$$\pi_{\text{ext},t}(a \mid s) \propto \exp\left[\frac{\beta_{\text{MB,ext}}}{\beta_{\text{MB,ext}} + \beta_{\text{MF,ext}}} Q^{(t)}_{\text{MB,ext}}(s,a) + \frac{\beta_{\text{MF,ext}}}{\beta_{\text{MB,ext}} + \beta_{\text{MF,ext}}} Q^{(t)}_{\text{MF,ext}}(s,a)\right]$$
$$\pi_{\text{int},t}(a \mid s) \propto \exp\left[\frac{\beta_{\text{MB,int}}}{\beta_{\text{MB,int}} + \beta_{\text{MF,int}}} Q^{(t)}_{\text{MB,int}}(s,a) + \frac{\beta_{\text{MF,int}}}{\beta_{\text{MB,int}} + \beta_{\text{MF,int}}} Q^{(t)}_{\text{MF, int}}(s,a)\right], \quad (8)$$

and as a result $\pi_t \propto \pi^{\beta_{\text{MB,ext}} + \beta_{\text{MF,ext}}}_{\text{ext},t} \cdot \pi^{\beta_{\text{MB,int}} + \beta_{\text{MF,int}}}_{\text{int},t}$.

In general, the contribution of seeking extrinsic reward and seeking intrinsic reward as well as the MB and MF branches to action-selection depends on different factors, including time passed since the beginning of the experiment[39,52], cognitive load[102], and whether the location of reward is known[18]. Here, we make a simplistic assumption that these contributions (expressed as the 4 inverse temperatures) are constant within but potentially different between the two phases of the experiment:

Phase 1: Before finding the goal state in episode 1 , we consider $\beta_{\mathrm{MB,ext}} = \beta^{(1)}_{\mathrm{MB,ext}}, \beta_{\mathrm{MF,ext}} = \beta^{(1)}_{\mathrm{MF,ext}}, \beta_{\mathrm{MB,\,int}} = \beta^{(1)}_{\mathrm{MB,int}}$, and $\beta_{\mathrm{MF,\,int}} = \beta^{(1)}_{\mathrm{MF,\,int}}$ as four independent free parameters chosen independently for each agent.

Phase 2: After finding the goal, i.e., in all episodes after episode 1, we consider $\beta_{\mathrm{MB,ext}} = \beta^{(2)}_{\mathrm{MB,\,ext}}, \beta_{\mathrm{MF,ext}} = \beta^{(2)}_{\mathrm{MF,\,ext}}, \beta_{\mathrm{MB,\,int}} = \beta^{(2)}_{\mathrm{MB,\,int}}$, and $\beta_{\mathrm{MF,\,int}} = \beta^{(2)}_{\mathrm{MF,\,int}}$ as another four independent free parameters chosen independently for each agent.

See 'Relative importance of novelty in action-selection' for how these inverse temperatures relate to the influence of intrinsic and extrinsic rewards on action-choices (Fig. 3.4.5C and Fig. 3.4.6).

Summary of free parameters: To summarize, the full hybrid algorithm has 22 free parameters:

$$\Phi = \{r_1^*, r_2^*, Q^{(0)}_{\mathrm{MF,ext}}, Q^{(0)}_{\mathrm{MF,\,int}}, \lambda_{\mathrm{ext}}, \lambda_{\mathrm{int}}, \mu_{\mathrm{ext}}, \mu_{\mathrm{int}}, \rho, \kappa, \epsilon_{\mathrm{new}}, \epsilon_{\mathrm{obs}}, T_{PS,\,\mathrm{ext}}, T_{PS,\,\mathrm{int}}, \beta^{(1)}_{\mathrm{MB,\,ext}}, \beta^{(2)}_{\mathrm{MB,\,ext}}, \beta^{(1)}_{\mathrm{MB,\,int}}, \beta^{(2)}_{\mathrm{MB,\,int}}, \beta^{(1)}_{\mathrm{MF,\,ext}}, \beta^{(2)}_{\mathrm{MF,\,ext}}, \beta^{(1)}_{\mathrm{MF,\,int}}, \beta^{(2)}_{\mathrm{MF,\,int}}\}, \quad (9)$$

where r_1^* and r_2^* are subjective values of the 3CHF goal and the 4CHF goal, respectively (with the 2 CHF goal being the reference goal with a value of 1), $Q^{(0)}_{\mathrm{MF,ext}}$ and $Q^{(0)}_{\mathrm{MF,int}}$ are the initial values for MF Q-values, λ_{ext} and λ_{int} are the discount factors, μ_{ext} and μ_{int} are the decay rates of the eligibility traces, ρ is the MF learning rate, κ is the leak parameter for model-building, ϵ_{new} and ϵ_{obs} are prior parameters for model-building, $T_{PS,\mathrm{ext}}$ and $T_{PS,\mathrm{int}}$ are the numbers of iterations for prioritized sweeping, and $\beta^{(1)}_{\mathrm{MB,ext}}, \beta^{(2)}_{\mathrm{MB,ext}}, \beta^{(1)}_{\mathrm{MB,\,int}}, \beta^{(2)}_{\mathrm{MB,\,int}}, \beta^{(1)}_{\mathrm{MF,ext}}, \beta^{(2)}_{\mathrm{MF,ext}}, \beta^{(1)}_{\mathrm{MF,\,int}}$, and $\beta^{(2)}_{\mathrm{MF,\,int}}$ are the inverse temperatures of the softmax policy.

Different choices of intrinsic reward

The intrinsic reward function $R_{\text{int},t}$ maps a transition $(s,a)\to s'$ to an intrinsic reward value, i.e., $r_{\text{int},t+1}=R_{\text{int},t}(s_t,a_t\to s_{t+1})$. In this section, we present our 3 choices of $R_{\text{int},t}$.

Novelty: For an agent seeking novelty (red in Fig. 3.4.2, Fig. 3.4.4, and Fig. 3.4.5), we define the intrinsic reward function as

$$R_{\text{int},t}(s,a\to s')=-\log p_N^{(t)}(s') \quad (10)$$

where $p_N^{(t)}(s')=\frac{1+\tilde{C}_{s'}^{(t)}}{1+|\mathcal{S}^{(t)}|+\sum_{s''}\tilde{C}_{s''}^{(t)}}$ is the state frequency with $\tilde{C}_{s'}^{(t)}$ the pseudo-count of encounters of state s' up to time t (similar to Eq. 5): $\tilde{C}_{s'}^{(t+1)}=\kappa\tilde{C}_{s'}^{(t)}+\delta_{s',s_{t+1}}$ with $\tilde{C}_{s'}^{(0)}=0$. With this definition, that generalizes earlier works[18] to the case where the number of states is unknown, the least novel states are those that have been encountered most often (i.e., with highest $\tilde{C}_{s'}^{(t)}$). Moreover, novelty is at its highest value for the unobserved states as we have $\tilde{C}_{s'}^{(t)}=0$ for any unobserved state $s'\notin\mathcal{S}^{(t)}$. Similar intrinsic rewards have been used in machine learning [19,20].

Surprise: For an agent seeking surprise (orange in Fig. 3.4.2, Fig. 3.4.4, and Fig. 3.4.5), we define the intrinsic reward function as the Shannon surprise (a.k.a. surprisal)[44]

$$R_{\text{int},t}(s,a\to s')=-\log p^{(t)}(s'\mid s,a), \quad (11)$$

where $p^{(t)}(s'\mid s,a)$ is defined in Eq. 6. With this definition, the expected (over s') intrinsic reward of taking action a at state s is equal to the entropy of the distribution $p^{(t)}(s'\mid s,a)$[103]. If $\epsilon_{\text{new}}<\epsilon_{\text{obs}}$, then the most surprising transitions are the ones to unobserved states. Similar intrinsic rewards have been used in machine learning[21,28].

Information-gain: For an agent seeking information-gain (green in Fig. 3.4.2, Fig. 3.4.4, and Fig. 3.4.5), we define the intrinsic reward function as

$$R_{\text{int},t}(s,a\to s')=\mathrm{D}_{\mathrm{KL}}\left[p^{(t)}(.\mid s,a)\parallel p^{(t+1)}(.\mid s,a)\right], \quad (12)$$

where D_{KL} is the Kullback-Leibler divergence[103], and $p^{(t+1)}$ is the updated world-model upon observing $(s,a)\to s'$. The dots in Eq. 12 denote the dummy variable over which we integrate to evaluate the Kullback-Leibler divergence. Note that if $s'\notin\mathcal{S}^{(t)}$, then there are some technical problems in the naïve computation of D_{KL} – since $p^{(t)}$ and $p^{(t+1)}$ have different supports. We deal with these problems using a more fundamental definition of D_{KL} using the Radon-Nikodym derivative; see Supplementary Materials for derivations and see ref. [42] for an alternative heuristic solution. Note that the information-gain in Eq. 12 has been also interpreted as a measure of surprise

(called 'Postdictive surprise'[92]), but it has a behavior radically different from that of the Shannon surprise introduced above (Eq. 11) - see ref. [44] for an elaborate treatment of the topic. Importantly, the expected (over s') information-gain corresponding to a state-action pair (s, a) converges to 0 as $\tilde{C}_{s,a}^{(t)} \to \infty$ (see Supplementary Materials for the proof). Similar intrinsic rewards have been used in machine learning[23,29,33,42].

Informing RL agents of different goal states and modeling optimism

Human participants had been informed that there were different goal states in the environment with different monetary reward values. This information was aimed to motivate participants to further explore the environment after they received the first reward at the end of episode one. This information is incorporated into our hybrid algorithms through a few mechanisms, where some include explicit information about the goal states but some others only an implicit notion of optimism.

Our main focus throughout the paper has been on modeling reward optimism by balancing intrinsic rewards against extrinsic rewards (Fig. 3.4.2, Fig. 3.4.5, and Fig. 3.4.6). In particular, assigning different values to $\beta_{\text{MB, ext}}$, $\beta_{\text{MF, ext}}$, $\beta_{\text{MB, int}}$, and $\beta_{\text{MF,int}}$ (c.f. Eq. 7) during the two phases of the experiment enables us to implicitly make the relative importance of intrinsic rewards depend on the difference between the reward value of the discovered goal r_{G^*} and the known reward values r_1^* and r_2^* of the other goal states (Eq. 2). Our results for the fitted relative importance of intrinsic reward across different groups of human participants (Fig. 3.4.6A) support this very assumption, which implies that the influence of reward optimism on the action-choices is via regulation of the balance between two separate policies, one for seeking intrinsic and one for seeking extrinsic rewards.

However, there are two other alternative mechanisms, purely based on seeking extrinsic rewards, that can contribute to reward-optimism in our hybrid algorithms: The model-based and model-free optimistic initialization. In this section, we discuss these mechanisms and how they balance exploration versus exploitation. We note that our results in Fig. 3.4.4 and Fig. 3.4.6 (particularly Fig. 3.4.6B) show that these two mechanisms alone are not enough and that a novelty-seeking module is necessary to explain the behavior of human participants; otherwise, all three intrinsically motivated algorithms would have the same probability of generating human data -

because the purely reward-seeking algorithm with optimistic initialization is a special case of all three intrinsically motivated algorithms that we compared. In other words, if optimistic initialization alone were sufficient to explain human behavior, then all three algorithms would perform equally well in Fig. 3.4.4 and the best fit would indicate a relative importance of 0 for novelty in Fig. 3.4.6.

Model-based optimistic initialization. MB optimistic initialization is an explicit approach to model reward-optimism through designing the world-model. The MB branch of the hybrid algorithm finds the extrinsic Q-values $Q^{(t)}_{\mathrm{MB,ext}}$ by solving the Bellman equations

$$Q^{(t)}_{\mathrm{MB,ext}}(s,a) = \bar{R}^{(t)}_{\mathrm{ext}}(s,a) + \lambda_{\mathrm{ext}} \sum_{s'} p^{(t)}(s' \mid s,a) \max_{a'} Q^{(t)}_{\mathrm{MB,ext}}(s',a') \quad (13)$$

where $p^{(t)}(s' \mid s,a)$ is estimated transition probability in Eq. 6, and

$$\begin{aligned} \bar{R}^{(t)}_{\mathrm{ext}}(s,a) &= \sum_{s'} p^{(t)}(s' \mid s,a) R_{\mathrm{ext}}(s,a \to s') \\ &= p^{(t)}\big(\tilde{G}_0 \mid s,a\big) + r_1^* p^{(t)}\big(\tilde{G}_1 \mid s,a\big) + r_2^* p^{(t)}\big(\tilde{G}_2 \mid s,a\big) \end{aligned} \quad (14)$$

is the average immediate extrinsic reward expected to be collected by taking action a in state s (see Eq. 2). Hence, the knowledge of the existence of three different goal states with three different rewards has an explicit influence on the MB branch of our algorithms. For example, because no transitions to any of the goal states have been experienced during episode 1, we have

$$\bar{R}^{(t)}_{\mathrm{ext}}(s,a) = \frac{\epsilon_{\mathrm{obs}}(1+r_1^*+r_2^*)}{\epsilon_{\mathrm{new}}+\epsilon_{\mathrm{obs}}|\mathcal{S}^{(t)}|+\tilde{C}^{(t)}_{s,a}} \quad (15)$$

This equation has two important implications. First, $\bar{R}^{(t)}_{\mathrm{ext}}(s,a)$ is an increasing function of ϵ_{obs}. This implies that the expected reward of a transition during episode 1 increases by increasing the prior probability of transition to states in $\mathcal{S}^{(t)}$. This is a direct consequence of our Bayesian approach to estimating the world-model. Second, $\bar{R}^{(t)}_{\mathrm{ext}}(s,a)$ is a decreasing function of $\tilde{C}^{(t)}_{s,a}$. This implies that the expected reward of a state-action pair decreases by experience. Importantly, $\bar{R}^{(t)}_{\mathrm{ext}}(s,a)$ converges to 0 as $\tilde{C}^{(t)}_{s,a} \to \infty$, which makes a link between exploration driven by the MB optimistic initialization and exploration driven by information-gain.

During episodes 2-5, the exact theoretical analysis of the MB optimistic initialization is rather complex. However, using a few approximation steps for episode 2, we can find a condition for whether the MB extrinsic Q-values show a preference for exploring or leaving the stochastic part (Supplementary Materials). The condition involves a comparison between the discounted

reward value of the discovered goal state $\lambda_{\text{ext}}^2 r_{G^*}$ and an optimistic estimate of a reward-to-be-found $R_{\text{Stoch.}}^{(t)}$ in the stochastic part that depends on $r_1^*, r_2^*, \lambda_{\text{ext}}, \epsilon_{\text{obs}}, \left|\mathcal{S}^{(t)}\right|$, and the average pseudo-count $\bar{C}^{(t)}$ of state-action pairs in the stochastic part (Supplementary Materials). We can show that if $r_{G^*} < r_2^*$, then increasing r_2^* would eventually result in a preference for staying in the stochastic part: If the reward value of a goal state is much greater than the value of the discovered goal state, then the agent prefers to keep exploring the stochastic part. However, for any value of r_2^* and r_{G^*}, increasing $\bar{C}^{(t)}$ would eventually result in a preference for leaving the stochastic part and going towards the already discovered goal: After a sufficiently long and unsuccessful exploration phase, agents will eventually give up exploration. This is another qualitative link between exploration based on the MB optimistic initialization and exploration driven by information-gain. This qualitative link leads to the conclusion that an agent with only the MB optimistic initialization cannot explain human behavior for the same reason that an agent with intrinsic reward based on information gain cannot explain human behavior.

Model-free optimistic initialization. As opposed to the MB branch of the hybrid algorithm, the MF branch does not have any explicit knowledge about the existence of different goal states and their values. However, the initial value $Q_{\text{MF,ext}}^{(0)}$ of the MF extrinsic Q-values quantifies an expectation of the reward values in the environment prior to any interaction with the environment. During episode 1, no extrinsic reward is received by the agent, hence, for a small enough learning rate ρ and an optimistic initialization $Q_{\text{MF,ext}}^{(0)} > 0$, the extrinsic reward prediction errors are always negative (Eq. 3). As a result, $Q_{\text{MF, ext}}^{(t)}(s, a)$ decreases as an agent keeps taking action a in state s, which motivates the agent to try new actions. This is a well-known mechanism for directed exploration in the machine learning community [37]. Similar to the MB optimistic initialization, the effect of the MF optimistic initialization fades out over time - which makes them both similar to exploration driven by information-gain.

During episode 2-5, the exact theoretical analysis of the MF optimistic initialization is complex and dependent on an agent's exact trajectory (because of the eligibility traces). However, whether the MF extrinsic Q-values show a preference for exploring or leaving the stochastic part essentially depends on the reward value of the discovered goal state r_{G^*} and the initialization value $Q_{\text{MF.ext}}^{(0)}$. For example, if an agent, starting at $s1$, takes the perfect trajectory of $s1 \rightarrow s2 \rightarrow s3 \rightarrow$

$s4 \rightarrow s5 \rightarrow s6 \rightarrow G^*$ in episode 1 , then, given a unit decay rate of the eligibility traces (i.e., $\mu_{\text{ext}} = 1$), it is easy to see that, in the 1st visit of state 4 in episode 2 , the agent prefers the stochastic/bad action over the progressing action if $r_{G^*} < \frac{1}{\lambda_{\text{ext}}^2}(1-\lambda_{\text{ext}})(1+\lambda_{\text{ext}}+\lambda_{\text{ext}}^2)Q_{\text{MF,ext}}^{(0)}$. This implies that, even though the MF branch is not explicitly aware of different goal states and their reward values, it is still able to model a type of reward optimism through initialization of Q-values. Nevertheless, since model fitting reveals an importance factor significantly greater than 0.5 (Fig. 3.4.6), the effective reward optimism generated by optimistic initialization is not strong enough to explain human behavior.

Efficient model-based planning for simulated participants

For simulating efficient agents in Fig. 3.4.2, we set $\varepsilon = 0$ (see Fig. 3.4.1A) and used a pure MB version of our algorithm with 13 parameters:

$$\left\{r_1^*, r_2^*, \lambda_{\text{ext}}, \lambda_{\text{int}}, \kappa, \epsilon_{\text{new}}, \epsilon_{\text{obs}}, T_{PS,\text{ext}}, T_{PS,\text{int}}, \beta_{\text{MB,ext}}^{(1)}, \beta_{\text{MB,ext}}^{(2)}, \beta_{\text{MB,int}}^{(1)}, \beta_{\text{MB,int}}^{(2)}\right\} \quad (16)$$

We considered perfect model-building by assuming $\kappa = 1$ and almost perfect planning by assuming $T_{PS,\text{ ext}} = T_{PS,\text{ int}} = 100$. We chose discount factors λ_{ext} and λ_{int} as well as prior parameters ϵ_{new} and ϵ_{obs} in the range of fitted parameters reported by Xu et al. (2021) [18]: $\lambda_{\text{ext}} = 0.95, \lambda_{\text{int}} = 0.70, \epsilon_{\text{new}} = 10^{-5}$ and $\epsilon_{\text{obs}} = 10^{-4}$. To relatively separate the effect of optimistic initialization [37] from seeking intrinsic reward in episode $2-5$, we assumed the same value of reward for all goals, i.e., $r_1^* = r_2^* = 1$. Finally, we considered $\beta_{\text{MB,ext}}^{(1)} = 0$ to have pure intrinsic reward seeking in episode 1 .

After fixing parameter values for 10 out of 13 parameters in Eq. 16, we fine-tuned $\beta_{\text{MB,int}}^{(1)}$ to minimize the average length of episode 1 (to find the goal as fast as possible; see Supplementary Materials). For episodes 2-5, we first set $\beta_{\text{MB, int}}^{(2)} = 0$ and $\beta_{\text{MB, ext}}^{(2)} = 10$ to have a non-deterministic policy for purely seeking extrinsic reward after the 1st encounter of the goal (the lightest shade of colors in Fig. 2). Different shades of color in Fig. 2 corresponds to different choices of $\omega \in [0,1]$ for $\beta_{\text{MB, int}}^{(2)} = \omega\beta_{\text{MB, int}}^{(1)}$ and $\beta_{\text{MB, ext}}^{(2)} = (1-\omega)\cdot 10$. More precisely, we used $\omega = 0$ for the darkest color (pure extrinsic reward seeking), $\omega = 1$ for the lightest color (pure intrinsic reward seeking), and $\omega = 0.7$ for the one in between. Higher values of ω indicates higher relative importance of the intrinsic reward.

Model-fitting and model-comparison

To compare seeking different intrinsic rewards based on their explanatory power, we considered our full hybrid (with both MF and MB components) algorithm - except that we put $T_{PS,\,\text{ext}} = T_{PS,\,\text{int}} = 100$ to decrease number of parameters, based on the results of Xu et al. (2021) [18] showing the negligible importance of this parameter. As a result, we had 20 free parameters for each of the three intrinsic rewards (i.e., novelty, information-gain, and surprise). For each intrinsic reward $R \in \{\text{novelty, inf-gain, surprise }\}$ and for each participant $n \in \{1, \dots, 57\}$, we estimated the algorithm's parameters by maximizing likelihood of data given parameters:

$$\hat{\Phi}_{n,R} = \arg\max_{\Phi} P(D_n \mid \Phi, R_n = R) \quad (17)$$

where D_n is the data of participant n, R_n is the intrinsic reward assigned to participant n, $P(D_n \mid \Phi, R_n = R)$ is the probability of D_n being generated by our intrinsically motivated algorithm seeking $R_n = R$ with its parameter equal to Φ (see Eq. 9), and $\hat{\Phi}_{n,R}$ is the set of estimated parameters that maximizes that probability. For optimization, we used Subplex algorithm[104] as implemented in Julia NLopt package[105].

Because all algorithms have the same number of parameters, we considered the maximum loglikelihood as the model log-evidence, i.e., for intrinsic reward R and participant n, we consider $\log P(D_n \mid R_n = R) \approx \log P\left(D_n \mid \hat{\Phi}_{n,R}, R_n = R\right)$ - which is equal to a shifted Schwarz approximation of the model log-evidence (also called BIC)[49,53]. Fig. 4A1 shows the total log-evidence $\sum_n \log P(D_n \mid R_n = R)$. With the fixed effects assumption at the level of models (i.e., assuming that $R_1 = R_2 = \cdots = R_{57} = R^*$), the total log-evidence is equal to the log posterior probability $\log P(R^* = R \mid D_{1:57})$ of R being the intrinsic reward used by all participants (plus a constant). See ref. [53,55] for tutorials.

We also considered the Bayesian model selection method of ref.[54] with the random effects assumption, i.e., assuming that participant n uses the intrinsic reward $R_n = R$, which is not necessarily the same as the one used by other participants, with probability P_R. We performed Markov Chain Monte Carlo sampling (using Metropolis Hasting algorithm[50] with uniform prior and 40 chains of length $10'000$) for inference and estimated the joint posterior distribution

$$P\left(R_{1:57}, P_{\text{novelty}}, P_{\text{inf-gain}}, P_{\text{surprise}} \mid D_{1:57}\right).$$

Fig. 3.4.4A2 shows the expected posterior probability $\mathbb{E}[P_R \mid D_{1:57}]$ as well as the protected exceedance probabilities $P(P_R > P_{R'} \text{ for all } R' \neq R \mid D_{1:57})$ computed by using the participant-wise log-evidences.

The boxplots of the fitted parameters of novelty-seeking are shown in Supplementary Materials. The same set of parameters were used for model-recovery in Fig. 3.4.4B, posterior predictive checks in Fig. 3.4.5, computing the relative importance of novelty in Fig. 3.4.6A-B (see 'Relative importance of novelty in action-selection'), and parameter recovery in Fig. 3.4.6C.

Posterior predictive checks, model-recovery, and parameter-recovery

For each intrinsic reward $R \in$ {novelty, inf-gain, surprise } and participant group $\mathcal{G} \in \{2\text{CHF}, 3\text{CHF}, 4\text{CHF}$, we repeated the following two steps 500 times: 1 . We sampled participant n from group $\mathcal{G}$ with probability $\frac{P(R_n=R|D_{1:57})}{\sum_{m\in\mathcal{G}}P(R_m=R|D_{1:57})}$. 2. We ran a 5 -episode simulations in our environment using the intrinsic reward R and the parameter $\hat{\Phi}_{n,R}$, i.e., we sampled a trajectory D from $P\left(D \mid \hat{\Phi}_{n,R}, R_n = R\right)$ (with the G^* of the environment corresponding to the group $\mathcal{G}$). As a result, we ended up with 1500 simulated participants (with randomly sampled parameters) for each algorithm.

We considered the simulated participants who took more than 3000 actions in any of the 5 episodes to be similar to the human participants who quit the experiment and excluded them from further analyses: $238(\sim 16\%)$ of simulated participants seeking novelty, $166(\sim 11\%)$ of those seeking information-gain, and $374(\sim 25\%)$ of those seeking surprise. We note that, even with the marginal influence of surprise on action-selection (Fig. 3.4.5C), one fourth of participants seeking surprise cannot escape the stochastic part in less than 3000 actions. Moreover, we excluded, separately for each algorithm, the simulated participants who took more than 3 times group-average number of actions in episodes 2-5 to finish the experiment (i.e., the same criterion that we used to detect nonattentive human participants): 45 (3%) of simulated participants seeking novelty, $77(\sim 5\%)$ of those seeking information-gain, and $27(\sim 2\%)$ of those seeking surprise. We then analyzed the remaining participants (1217 simulated participants seeking novelty, 1257 seeking informationgain, and 1099 seeking surprise) as if they were real human participants. Fig. 3.4.5 and

its supplements in Supplementary Materials show the data statistics of simulated participants in comparison to human participants.

Given the participants simulated by each of the three intrinsically motivated algorithms, we fitted all three algorithms to the action-choices of 150 simulated participants (50 from each participant group, i.e., 2CHF, 3CHF, and 4CHF). Then, we applied the Bayesian model selection method of ref.[54] to 5 randomly chosen sub-populations of these 150 simulated participants (each with 60 participants, i.e., 20 from each participant group). Fig. 3.4.4B shows the results of the model comparison averaged over these 5 repetitions. Fig. 3.4.6C shows the relative importance of novelty in action-selection (see Eq. 18) for each of the 150 simulated participants estimated using the original parameters (which were used for simulations) and the recovered parameters (which were found by re-fitting the algorithms to the simulated data).

Relative importance of novelty in action-selection

The relative importance of novelty in action-selection depends not only on the inverse-temperatures $\beta_{\mathrm{MB,\,ext}}$, $\beta_{\mathrm{MF,\,ext}}$, $\beta_{\mathrm{MB,\,int}}$, and $\beta_{\mathrm{MF,int}}$ but also on the variability of Q-values; for example, if the extrinsic Q-values $Q^{(t)}_{\mathrm{MB,\,ext}}(s,a)$ and $Q^{(t)}_{\mathrm{MF,\,ext}}(s,a)$ are the same for all state-action pairs, then, independently of the values of the inverse-temperatures, the action is taken by a pure novelty-seeking policy - because the policy in Eq. 7 can be re-written as $\pi^{(t)}(a \mid s) \propto \exp\left[\beta_{\mathrm{MB,\,int}}\, Q^{(t)}_{\mathrm{MB,\,int}}(s,a) + \beta_{\mathrm{MF,\,int}}\, Q^{(t)}_{\mathrm{MF,\,int}}(s,a)\right]$. Thus, to measure the contribution of different components of action-selection to the final policy, we need to consider the variations in their Q-values as well.

In this section, we propose a variable $\omega_{\mathrm{i2e}} \in [0,1]$ for quantifying the relative importance of seeking intrinsic reward in comparison to seeking extrinsic reward. We first define total intrinsic and extrinsic Q -values as $Q^{(t)}_{\mathrm{ext}}(s,a) = \beta_{\mathrm{MB,ext}} Q^{(t)}_{\mathrm{MB,ext}}(s,a) + \beta_{\mathrm{MF,ext}} Q^{(t)}_{\mathrm{MF,ext}}(s,a)$ and $Q^{(t)}_{\mathrm{int}}(s,a) = \beta_{\mathrm{MB,\,int}}\, Q^{(t)}_{\mathrm{MB,\,int}}(s,a) + \beta_{\mathrm{MF,\,int}}\, Q^{(t)}_{\mathrm{MF,\,int}}(s,a)$, respectively. We further define the state-dependent variations in Q -values as $\Delta Q^{(t)}_{\mathrm{ext}}(s) = \max_a Q^{(t)}_{\mathrm{ext}}(s,a) - \min_a Q^{(t)}_{\mathrm{ext}}(s,a)$ and $\Delta Q^{(t)}_{\mathrm{int}}(s) = \max_a Q^{(t)}_{\mathrm{int}}(s,a) - \min_a Q^{(t)}_{\mathrm{int}}(s,a)$ as well as their temporal average $\Delta \bar{Q}_{\mathrm{ext}} = \left\langle \Delta Q^{(t)}_{\mathrm{ext}}(s_t) \right\rangle$ and $\Delta \bar{Q}_{\mathrm{int}} = \left\langle \Delta Q^{(t)}_{\mathrm{int}}(s_t) \right\rangle$, where $\langle . \rangle shows the temporal average.$ $\Delta \bar{Q}_{\mathrm{ext}}$ and $\Delta \bar{Q}_{\mathrm{int}}$ show the average difference between the most and least preferred action with respect to seeking

extrinsic and intrinsic reward, respectively. Therefore, a feasible way to measure the influence of seeking intrinsic reward on action-selection is to define ω_{i2e} as

$$\omega_{\mathrm{i2e}} = \frac{\Delta\bar{Q}_{\mathrm{int}}}{\Delta\bar{Q}_{\mathrm{ext}} + \Delta\bar{Q}_{\mathrm{int}}}$$

Fig. 3.4.6 A shows the value ω_{i2e} in episode $2-5$ computed for each human participant (dots) and averaged over different groups (bars), and Fig. 3.4.6B shows the same for episode 1. Fig. 3.4.5C shows the value ω_{i2e} in episode $2-5$ for the 2 CHF group of simulated participants. See Supplementary Materials for a similar approach for quantifying the relative importance of the MB and MF policies in action-selection.

3.4 *Results*

Experimental Paradigm

We first design an experimental paradigm for human participants that allows us dissociate predictions of different intrinsically motivated RL algorithms. We employ a sequential decision-making paradigm[34–36] for navigation in an environment with 58 states plus three goal states (Fig. 3.4.1A-B). Three actions are available in each non-goal state, and agents can move from one state to another by choosing these actions (arrows in Fig. 3.4.1A-B). We use the term 'agents' to refer to either human participants or agents simulated by RL algorithms. In the human experiments, states are represented by images on a computer screen and actions by three disks below each image (Fig. 1C); for simulated participants, both states and actions are abstract entities (i.e., we consider RL in a tabular setting[37]). The assignment of images to states and disks to actions is random but fixed throughout the experiment. Agents are informed that there are three different goal states in the environment (G^*, G1, or G2 in Fig. 3.4.1A) and that their task is to find a goal state 5 times; see Methods for how this information is incorporated in the RL algorithms. Importantly, neither human participants nor RL agents are aware of the total number of states or the structure of the environment (i.e., how states are connected to each other).

The 58 states of the environment can be classified into three groups: Progressing states (1 to 6 in Fig. 3.4.1A), trap states (7 and 8 in Fig. 3.4.1A), and stochastic states (S-1 to S-50 in Fig. 3.4.1B, shown as a dashed oval in Fig. 3.4.1A). In each progressing state, one action ('progressing' action) takes agents one step closer to the goals and another action ('bad' action) takes them to one of the trap states. The third action in states 1-3 and 5-6 is a 'self-looping' action that makes agents stay at the same state. Except for the progressing action in state 6, all these actions are deterministic, meaning that they always lead to the same next state. The progressing action in state 6 is almost deterministic: It takes participants to the 'likely' goal state G^* with a probability of $1 - \varepsilon$ and to the 'unlikely' goal states G1 and G2 with equal probabilities of $\frac{\varepsilon}{2} \ll 1$. In state 4, instead of a self-looping action, there is a 'stochastic' action that takes agents to a randomly chosen (with equal probability) stochastic state (Fig. 3.4.1B1). In each stochastic state, one action takes agents back to state 4 and two stochastic actions take them to another randomly chosen stochastic state (Fig. 1B2). In each trap state, all three actions are deterministic: Two actions bring agents to either the same or the other trap state and one action to state 1.

The stochastic part of the environment – which is inspired by the machine-learning literature and mimics the main features of a 'noisy TV'[28] – is a crucial difference to existing paradigms in the literature of behavioral neuroscience[18,38,39]. Without the stochastic part, intrinsic motivation helps agents to avoid the trap states and find the goal[18], hence it helps exploration before and does not harm exploitation after finding a goal. By adding the stochastic part, we aim to quantify how much exploitation of the discovered goal is reduced because of the distraction by the stochastic states.

We organize the experiment in 5 episodes: Agents are randomly initialized at state 1 or 2 and are instructed to find a goal 5 times. After finding a goal, agents are randomly re-initialized at state 1 or 2. We choose a small enough ε (Fig. 3.4.1A) to safely assume that all agents visit only G^* while being aware that G1 and G2 exist (Methods).

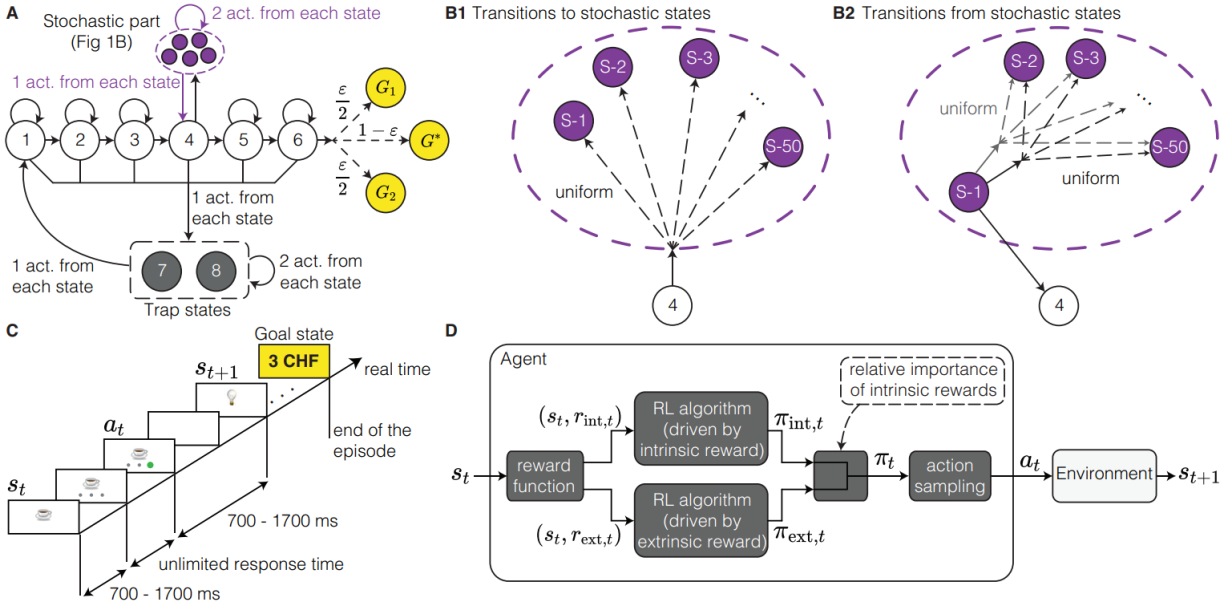

Figure 3.4.1: Experimental paradigm and computational model. A. Structure of the environment with the stochastic states merged together (dashed oval; see B). Each circle represents a state and each solid arrow an action. All actions except for the ones to the stochastic part or to the goal states are deterministic. Dashed arrows indicate random transitions; values (e.g., $1-\varepsilon$) show probabilities of each transition. We choose $\varepsilon \ll 1$ (see Methods). **B.** Structure of the stochastic part of the environment (states S-1 to S-50), i.e., the dashed oval in A. **B1.** In state 4, one action takes agents randomly (with uniform distribution) to one of the stochastic states. **B2.** In each stochastic state (e.g., state S-1 in the figure), one action takes agents back to state 4 and two actions to another randomly chosen stochastic state. **C.** Time-line of one episode in human experiments. The states are represented by images on a computer screen and actions by disks below each image. An episode ends when a goal image (i.e., '3 CHF' image in this example) is found. **D.** Block diagram of the intrinsically motivated RL algorithm. Given the state s_t at time t, the intrinsic reward $r_{\text{int},t}$ (i.e., novelty, information-gain, or surprise) and the extrinsic reward $r_{\text{ext},t}$ (i.e., the monetary reward value of s_t) are evaluated by a reward function and passed to two identical (except

for the reward signals) parallel RL algorithms. The two algorithms compute two policies, one for seeking intrinsic reward $\pi_{\text{int},t}$ and one for seeking extrinsic reward $\pi_{\text{ext},t}$. The two policies are then weighted according to the relative importance of the intrinsic reward and are combined to make a single hybrid policy π_t. The next action a_t is selected by sampling from π_t. See Methods for details.

Simulating intrinsically motivated agents with efficient algorithms

To formulate qualitative predictions for human behavior, we simulate three intrinsically motivated RL algorithms. Intrinsic motivation is described in each algorithm by 'intrinsic rewards' that agents give to themselves upon visiting 'novel', 'surprising', or 'informative' states (see Methods for details). Extrinsic rewards, on the other hand, are received only when visiting the three goal states. Agents simulated by each algorithm are able to navigate in an environment with an unknown number of states by seeking a combination of extrinsic and intrinsic rewards (Fig. 3.4.1D): At each time t, an agent observes state s_t and evaluates an extrinsic reward value $r_{\text{ext},t}$ (which is zero except at the goal states) and an intrinsic reward value $r_{\text{int},t}$ (e.g., novelty of state s_t). Extrinsic and intrinsic reward values are then passed to two parallel blocks of RL, each working with a single reward signal. Independently of each other, the two blocks use efficient model-based planning[40,41] to propose a policy $\pi_{\text{ext},t}$ that maximizes future extrinsic rewards and $\pi_{\text{int},t}$ that maximizes future intrinsic rewards[18,26], respectively. The two policies are combined into a hybrid policy πt for taking the next action at, controlled by a set of free parameters that indicate the relative importance of intrinsic over extrinsic rewards (Methods). The degree of exploration is high if $\pi_{\text{int},t}$ dominates $\pi_{\text{ext},t}$ during action-selection.

For the intrinsic reward $r_{\text{int},t}$, we choose one option from each of the three main categories of intrinsic rewards in machine learning[26,27]: (i) novelty[18–20] quantifies how infrequent the state s_t has been until time t; (ii) information-gain[23,25,42,43] quantifies how much the agent updates its belief about the structure of the environment upon observing the transition from the state-action pair (s_{t-1}, a_{t-1}) to state s_t; and (iii) surprise[21,28,44] quantifies how unexpected it is to observe state s_t after taking action a_{t-1} at state s_{t-1}.

The three different intrinsic reward signals lead to three efficient intrinsically motivated algorithms and to three groups of simulated efficient agents: those (i) seeking novelty, (ii) seeking information-gain, and (iii) seeking surprise. In the following section, we focus on episodes 2-5 and

formulate the qualitative predictions of different intrinsically motivated RL algorithms for human behavior by characterizing the behavior of these simulated efficient agents. We note that these predictions are made by using efficient RL algorithms with perfect memory and high computational power. Thus these efficient agents are a starting point and used only (i) to test whether our experimental paradigm dissociates action-choices of different intrinsically motivated RL algorithms and (ii) to gain insights about their principal differences; a more realistic simulation of human behavior is presented in a later section.

Different intrinsically motivated algorithms exhibit principally different behavioral patterns

To avoid arbitrariness in the choice of parameters, we fine-tune the parameters of each algorithm to have on average the lowest number of actions in episode 1 (to have the most efficient exploration; Methods). As a result, different algorithms achieve a similar performance during episode 1 and find the goal $G*$ almost equally fast (Supplementary Materials). Hence, exploration policies driven by different intrinsic rewards cannot be qualitatively distinguished during episode 1.

Given the same set of parameters, we study how different simulated efficient agents behave in episodes 2-5 (Fig. 3.4.2). After finding the goal G^* for the 1st time, an agent has two options: (i) return to the discovered goal state G^* (exploitation) or (ii) search for the other goal states G1 and G2 (exploration). In our simulations, we consider three choices for the trade-off between exploration and exploitation by changing the relative importance of $\pi_{int,t}$ over $\pi_{ext,t}$ (Fig. 3.4.1D): pure exploitation (the action policy does not depend on intrinsic rewards, i.e., $\pi_t = \pi_{ext,t}$), pure exploration (the action policy does not depend on extrinsic rewards, i.e., $\pi_t = \pi_{int,t}$), and a mixture of both (different shades of each color in Fig. 3.4.2). If the extrinsic reward value assigned to G1 or G2 is higher than the one assigned to G^*, then the policy $\pi_{ext,t}$ for seeking extrinsic rewards can also contribute to exploration in episodes 2-5 (Methods). In order to characterize qualitative features essential to exploration driven by different intrinsic rewards, we assume a symmetry between the three goal states in the simulated efficient agents and assign the same extrinsic reward value to all goals (Methods); we drop this assumption in the next sections and quantify the additional but negligible contribution of $\pi_{ext,t}$ to explaining human exploration.

For all three groups of simulated efficient agents, decreasing the relative importance of intrinsic rewards decreases both the search duration (Fig. 3.4.2A1, C1, and E1) and the fraction of time spent in the stochastic part (Fig. 3.4.2A2, C2, and E2). This observation implies that intrinsically motivated exploration leads to an attraction to the stochastic part of the environment, effectively keeping the simulated efficient agents away from the goal region beyond state 6 (Fig. 3.4.1A). Our results thus confirm earlier findings in machine learning[26,28] that intrinsically motivated agents get distracted by noisy reward-independent stimuli.

While all three groups of simulated efficient agents get distracted by the stochastic part, their degree of distraction is different (different colors in Fig. 3.4.2A3, C3, and E3). For efficient agents that purely seek information-gain (i.e., pure exploration), the time spent in the stochastic part decreases over episodes (Fig. 3.4.2C3), whereas we observe the opposite pattern for efficient agents that purely seek novelty (Fig. 3.4.2A3) or surprise (Fig. 3.4.2E3). In particular, efficient agents that purely seek surprise get most often (i.e., in > 50% of simulations in episode 5) stuck in the stochastic part and do not escape it within 3000 actions (Fig. 3.4.2E3). These observations confirm the inefficiency of seeking surprise and the efficiency of seeking information-gain in dealing with noise[26].

In order to further dissociate action-choices of different algorithms, we analyze the action preferences of simulated efficient agents in state 4 during episode 2 (Fig. 3.4.2B, D, and F). For all three

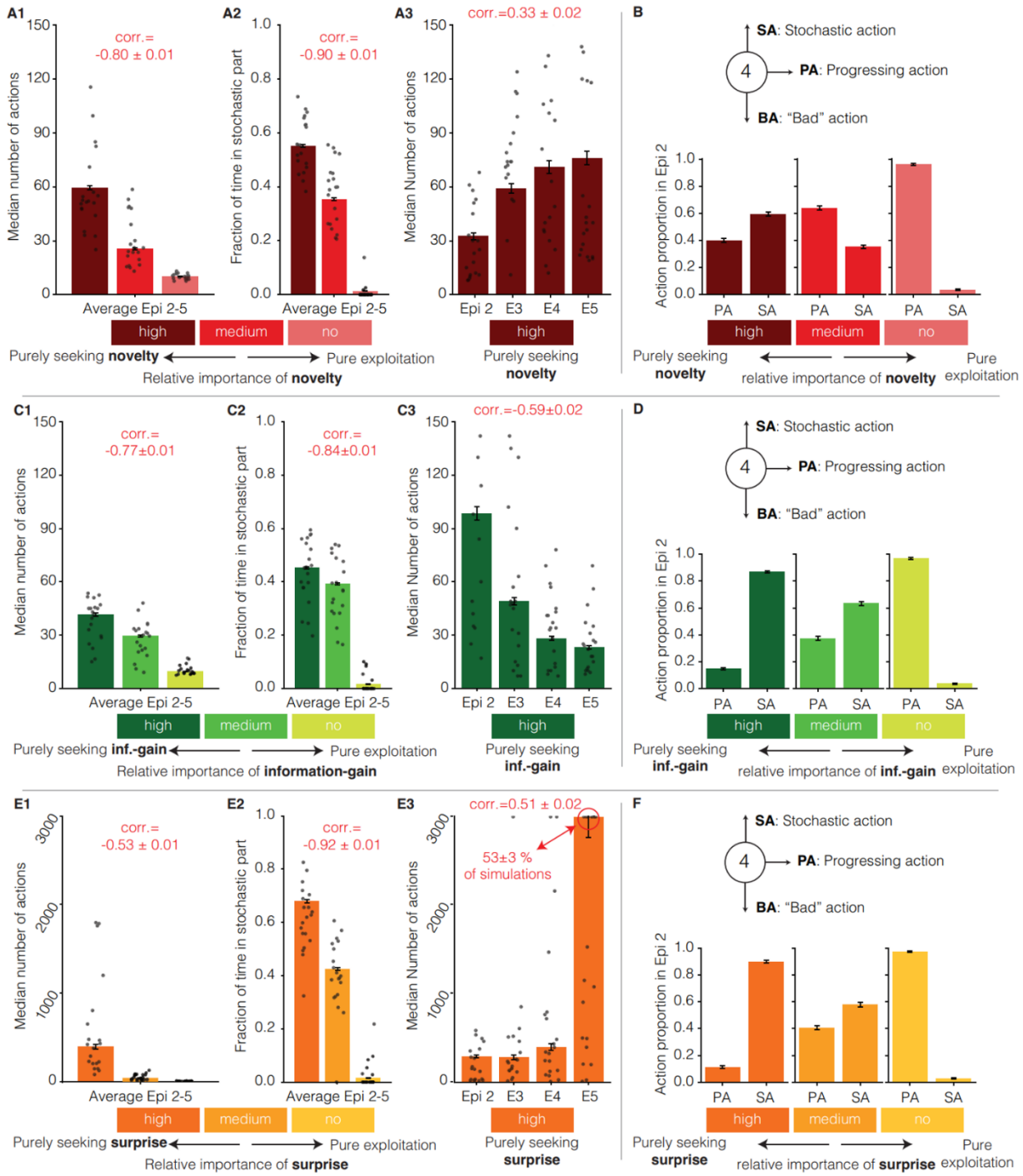

Figure 3.4.2: Qualitative predictions for episodes 2-5 with efficient algorithms seeking novelty (A-B), surprise (C-D), and information-gain (E-F) as intrinsic rewards. We consider three levels of importance for intrinsic rewards (Fig. 3.4.1D): high (dark colors), medium (shaded colors), and no (light colors). For each level, we run 500 simulations of each algorithm. **A1, C1, and E1.** Median number of actions over episodes 2-5. Error bars show the standard error of the median (SEMed; evaluated by bootstrapping). Single dots show the data of 20 (randomly chosen out of 500) individual simulations to illustrate variabilities among simulations. Simulations stopped after 3000 actions even if a goal state was

not reached. The Pearson correlation between the search duration and the degree of exploitation is negative (red numbers), indicating that search duration decreases if the degree of exploitation increases (Methods). **A2, C2, and E2.** Average fraction of time spent in the stochastic part of the environment during episodes 2-5. The Pearson correlation between the fraction of time spent in the stochastic part and the degree of exploitation is negative (Methods). Error bars show the standard error of the mean (SEMean) and single dots the data of 20 individual simulations. **A3, C3, and E3.** Median number of actions in episodes 2-5 for simulated efficient agents *purely* driven by intrinsic rewards (i.e., pure exploration). The Pearson correlation between the search duration and episode number is positive for seeking novelty or surprise but is negative for seeking information-gain. Error bars show the SEMed and single dots the data of 20 individual simulations. **B, D, and F.** Fraction of time taking the progressing action (PA) and the stochastic action (SA) when encountering state 4 during episode 2. Purely seeking novelty shows a smaller difference between the preference for SA and PA in state 4 compared to purely seeking information-gain or surprise. Error bars show the SEMean.

groups of efficient agents, increasing the relative importance of intrinsic rewards increases their preference for the stochastic action. However, for the highest importance of intrinsic rewards, the probability of choosing the progressing action is substantially lower than the probability of choosing the stochastic action for seeking surprise or information-gain (15% vs. 85%; Fig. 3.4.2D and F), whereas this difference is much smaller for seeking novelty (40% vs. 60%; Fig. 3.4.2B). This distinct behavior of novelty-seeking is due to the fact that novelty is defined for states, whereas surprise and information-gain are defined for transitions (i.e., state-action pairs; see Methods): By the end of episode 1, the goal state has been observed only once and remains, during episode 2, relatively novel (and hence attractive for an efficient novelty-seeking agent) compared to most stochastic states, whereas there are many actions between the stochastic states that have rarely or potentially never been chosen and are, thus, attractive for an efficient agent seeking surprise or information-gain.

To summarize, different intrinsically motivated algorithms exhibit principally different behavioral patterns in our experimental paradigm. We consider these behavioral patterns as qualitative predictions for human behavior.

Human participants

To test the predictions of intrinsically motivated algorithms, we first compare the exploratory behavior of human participants with that of simulated efficient agents. For simulated efficient agents, the relative importance of the intrinsic reward for action-selection (Fig. 3.4.1D) determines the balance of exploration versus exploitation. A challenge in human experiments is that we do not have explicit control over the variable that controls the relative importance of

intrinsic rewards compared to extrinsic rewards. Inspired by earlier studies[45–47], we conjecture that human participants who are more optimistic about finding a goal with a high value of reward are more curious to explore the environment than human participants who are less optimistic. In other words, we hypothesize that the relative importance of intrinsic rewards in human participants is positively correlated with their degree of 'reward optimism', where we define reward optimism as the expectancy of finding a goal of higher value than those already discovered.

Based on this hypothesis, we include a novel reward manipulation in the instructions given at the beginning of the experiment: We inform human participants that there are three different possible reward states corresponding to values of 2 Swiss Franc (CHF), 3 CHF, and 4 CHF, represented by three different images (Methods). At the beginning of the experiment, we randomly assign the three different reward values to the goal states G^*, G1, and G2 in Fig. 3.4.1A, separately for each participant (without informing them), and keep the assignment fixed throughout the experiment. After this random assignment, G^* has a different value for different participants. Even though all participants receive the same instructions, participants who are randomly assigned to an environment with 4 CHF reward value for G^* do not have any monetary incentive to further explore in episodes 2-5 (= a low degree of reward optimism), whereas participants who are assigned to an environment with 2 CHF reward value for G^* are likely to keep searching for more valuable goals in episodes 2-5 (= a high degree of reward optimism). Therefore, we have three different groups of participants with three different levels of reward optimism in episodes 2-5; see Methods for how this information is incorporated in the RL algorithms. We note that our definition of reward optimism in the context of our experiment is in line but independent of the notion of general optimism that is quantified for individual participants in psychology[48].

Following a power analysis based on the data of simulated efficient agents (Methods), we recruited 63 human participants and collected their action-choices during the 5 episodes of our experiment: 23 participants in an environment with 2 CHF reward value for G^* and two times 20 human participants in environments with 3 CHF and 4 CHF reward value for G^*, respectively. In the rest of the manuscript, we refer to each group by their reward value of G^*, e.g., the 3 CHF group is the group of human participants who were assigned to 3 CHF reward value for G^* (as in Fig. 3.4.1C). We excluded the data of 6 human participants from further analyses since they either did not finish the experiment or had an abnormal performance (Methods).

Human participants exhibit a persistent distraction by stochasticity

We perform the same series of analyses on the behavior of human participants as those performed on the behavior of simulated efficient agents (Fig. 3.4.3). In episodes 2-5, the search duration of human participants (Fig. 3.4.3A1) and the fraction of time they spend in the stochastic part (Fig. 3.4.3A2) are both negatively correlated with the goal value of their environment, e.g., the 2 CHF group has a longer search duration and spends more time in the stochastic part than the other two groups. Moreover, increasing the goal value increases the preference of human participants for the progressing action in state 4 during episode 2 (Fig. 3.4.3B). These observations support our hypothesis that increasing the degree of reward optimism influences the behavior of human participants in the same way as increasing the relative importance of intrinsic rewards influences the behavior of simulated efficient agents (e.g., compare Fig. 3.4.3A1, A2, and B with Fig. 3.4.2A1, A2, and B, respectively).

The behavior of the 2 CHF group is particularly interesting since they are the most optimistic group of participants. The 2 CHF group exhibits a constant search duration over episodes 2-5 (zero correlation accepted by Bayesian hypothesis testing[49]; Fig. 3.4.3A3). This implies that they

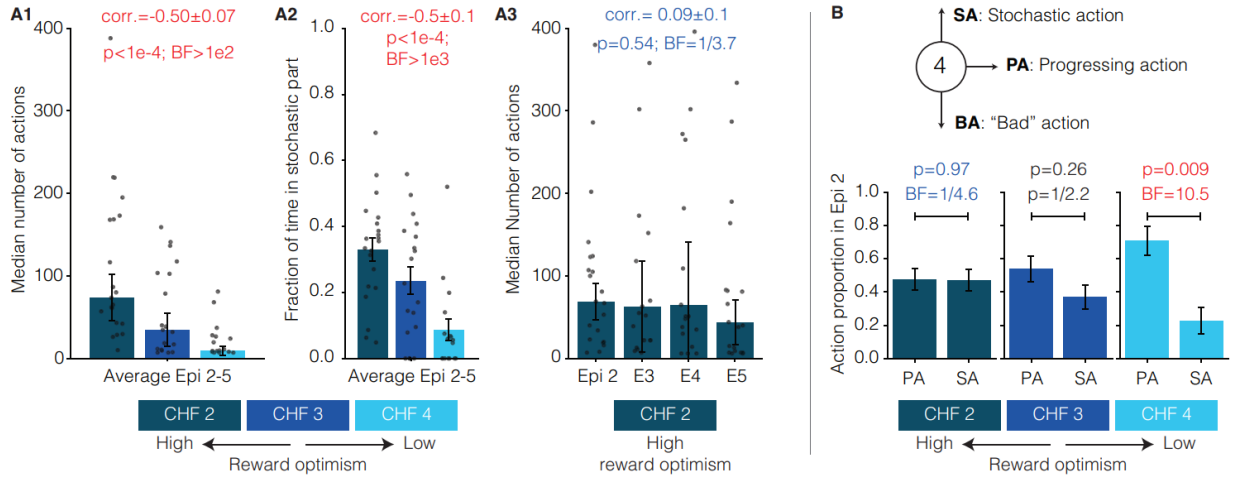

Figure 3.4.3: Human participants persistently explore the stochastic part if they are highly optimistic. A. Search duration in episodes 2-5. **A1.** Median number of actions over episodes 2-5 for the three different groups: 2 CHF (dark), 3 CHF (medium), and 4 CHF (light). Error bars show the SEMed (evaluated by bootstrapping) and single dots the data of individual participants. The Pearson correlation between the search duration and the goal value is negative (correlation test; t = −4.2; 95% Confidence Interval (CI) = (−0.67, −0.27); Degree of Freedom (DF) = 55; Methods). **A2.** Average fraction of time spent in the

stochastic part of the environment during episodes 2-5. The Pearson correlation between the fraction of time spent in the stochastic part and the goal value is negative (correlation test; t = −4.7; 95% CI = (−0.70, −0.32); DF = 55; Methods). Error bars show the SEMean and single dots the data of individual participants. **A3.** Median number of actions in episodes 2-5 for the 2 CHF group. A Bayes Factor (BF) of 1/3.7 in favor of the null hypothesis 49 suggests a zero Pearson correlation between the search duration and the episode number (one-sample t-test on individual correlations; t = 0.63; 95% CI = (−0.20, 0.37); DF = 20). Error bars show the SEMed and single dots the data of individual participants. **B.** Fraction of time choosing the progressing action (PA) and the stochastic action (SA) when encountering state 4 during episode 2; see Supplementary Materials for other progressing states. Error bars show the SEMean. The difference between PA and SA for the 4 CHF group is significant (one-sample t-test; t = 2.99; 95% CI = (0.14, 0.81); DF = 16). A BF of 1/4.6 in favor of the null hypothesis[49] suggests an equal average between PA and SA for the 2 CHF group (one-sample t-test; t = 0.039; 95% CI = (−0.25, 0.26); DF = 20). The test for 3 CHF group is inconclusive (one-sample t-test; t = 1.17; 95% CI = (−0.13, 0.47); DF = 18). Red p-values: Significant effects with False Discovery Rate controlled at 0.0550 (see Methods). Red BFs: Significant evidence in favor of the alternative hypothesis (BF≥ 3). Blue BFs: Significant evidence in favor of the null hypothesis (BF≤ 1/3).

persistently explore the stochastic part. Moreover, during episode 2, the 2 CHF group chooses the progressing and the stochastic actions equally often (no-difference in means accepted by Bayesian hypothesis testing[49]; Fig. 3.4.3B). If we assume that the high degree of reward optimism in the 2 CHF group results in a policy that is driven dominantly by intrinsic rewards (driving exploration) and only marginally by extrinsic rewards, then these observations are more similar to the qualitative predictions of seeking novelty than those of seeking information-gain or surprise (compare Fig. 3.4.3B against Fig. 3.4.2B, D, and F).

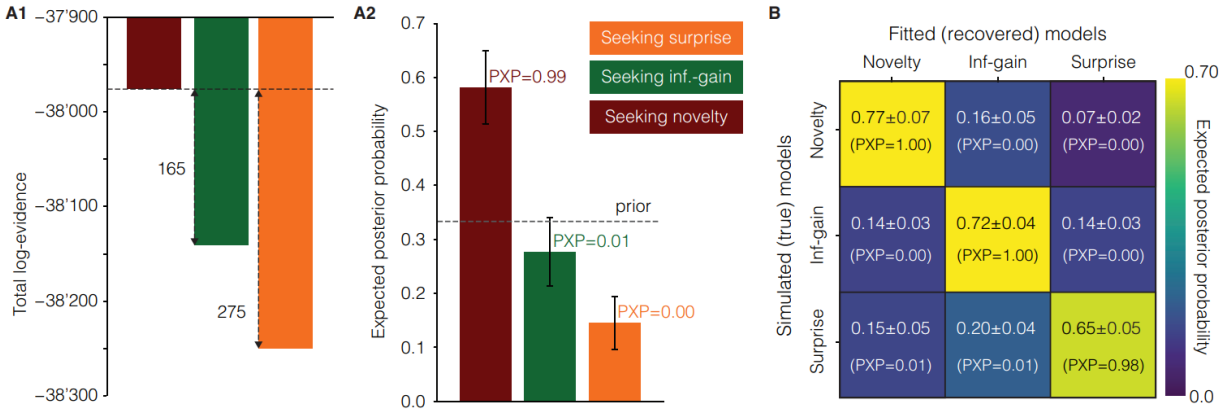

Figure 3.4.4: Novelty-seeking is the most probable model of human behavior. A. Human participants' action-choices are best explained by novelty-seeking (see Methods for details). **A1.** Model log-evidence summed over all participants (i.e., assuming that different participants have the same exploration strategy but can have different parameters[53]) is significantly higher for seeking novelty than seeking information-gain or surprise. High values indicate good performance, and differences greater than 10 are traditionally[50] considered as strongly significant. **A2.** The expected posterior model probability with random effects assumption[54] (i.e., assuming that different participants can have different exploration strategies and different parameters) given the data of all participants. PXP stands for Protected Exceedance Probability[54], i.e., the probability of one model being more probable than the others. Error bars show the standard deviation of the posterior distribution. **B.** Confusion matrix from the model recovery procedure: Each row shows the results of applying our model-fitting and -comparison procedure (as in **A2**) to the action-choices of simulated participants by one of the three algorithms (with their parameters fitted to human data; see Methods). Color-code shows the expected posterior probability and numbers in parentheses the PXP (both averaged over 5 sets of 60 simulated participants). We could always recover the model that had generated the data (PXP $\geq$ 0.98), using almost the same number of simulated participants (60) as human participants (57).

Novelty-seeking is the most probable model of human exploration

In the previous section, we observed that human participants exhibit patterns of behavior qualitatively similar to those of novelty-seeking simulated efficient agents. However, the qualitative predictions in Fig. 3.4.2 were made based on the assumptions of (i) using efficient RL algorithms with perfect memory and high computational power, (ii) using parameters that were optimized for the best performance in episode 1, and (iii) assigning the same extrinsic reward value to different goal states. In this section, we use a more realistic model of behavior than that of efficient agents in Fig. 3.4.2: In order to model the behavior of human participants, we use a hybrid RL model[18,36,38,51] combining model-based planning[41] and model-free habit-formation[52], account for imperfect memory and suboptimal choice of parameters, and allow our algorithms to assign

different extrinsic reward values to different goal states (Methods). We fit the parameters of our three intrinsically motivated algorithms to the action-choices of each individual participant by maximizing the likelihood of data given parameters (Methods). Such a flexible modeling approach allows each of the three algorithms to find its closest version to the behavior of human participants, constrained on using one specific intrinsic reward signal (i.e., novelty, surprise, or information-gain).

Given the fitted algorithms, we use Bayesian model-comparison[53,54] to quantitatively test whether human behavior is explained better by seeking novelty than seeking information-gain or surprise

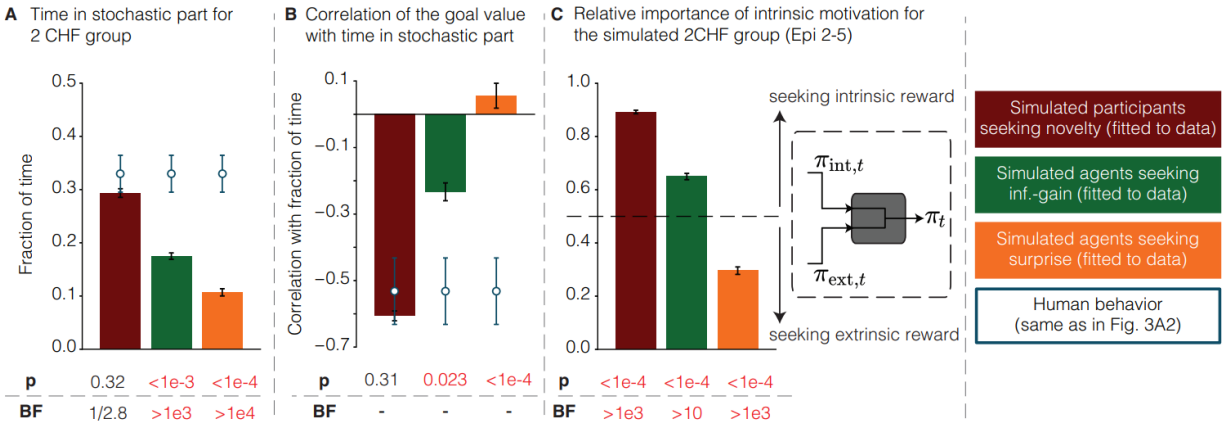

Figure 3.4.5: Seeking novelty, but not surprise or information-gain, can reproduce data statistics. For each of the three intrinsic rewards, we run 1500 simulations of algorithms with parameters fitted to individual human participants; random seeds are different in each simulation. We divide the simulated participants into three groups (corresponding to the 2 CHF, 3 CHF, and 4 CHF goal values) and use the same criteria as we used for human participants to detect and remove outliers among simulated participants (Methods). **A.** Average fraction of time during episodes 2-5 spent by the 2 CHF group of human participants (blue circles, same data as in Fig. **3A2**) and the simulated participants (bars). Error bars: SEMean. P-value and BF: Comparison between the simulated and human participants (unequal variances t-test). Human participants spend a significantly greater fraction of their time in the stochastic part than simulated participants seeking information-gain (t = 4.4; 95% CI = (0.08, 0.23); DF = 21.2) or surprise (t = 6.3; 95% CI = (0.15, 0.30); DF = 21.6). No significant difference was observed for novelty seeking (t = 1.0; 95% CI = (−0.04, 0.11); DF = 22.3). **B.** Pearson correlation between the fraction of time during episodes 2-5 spent in the stochastic part and the goal value. Human participants' data shows the same correlation value as reported in Fig. **3A2**. Error bars: Standard deviation evaluated by bootstrapping. P-values are from permutation tests (1000 sampled permutations; Bayesian testing was not applicable). **C.** The relative contribution of intrinsic rewards (i.e., dominance of $\pi_{int,t}$ over $\pi_{ext,t}$; Eq. 18 in Methods) in episodes 2-5 for the 2 CHF group of simulated participants. P-value and BF: Comparison with 0.5 (one-sample t-test). We observe a dominance of $\pi_{int,t}$ for seeking novelty (t = 58.2; 95% CI = (0.88, 0.91); DF = 416) and information-gain (t = 12.7; 95% CI = (0.63, 0.67); DF = 379) but a dominance of $\pi_{ext,t}$ for seeking surprise (t = −14.6; 95% CI = (0.27, 0.32); DF = 327). Red p-values: Significant effects with False Discovery Rate

controlled at 0.05[50] (see Methods). Red BFs: Significant evidence in favor of the alternative hypothesis (BF≥ 3).

(Methods). Our model-comparison results show that seeking novelty is the most probable model for the majority of human participants, followed by seeking information-gain as the 2nd most probable model (Fig. 3.4.4A; Protected Exceedance Probability[54] = 0.99 and 0.01 for seeking novelty and information-gain, respectively). This result shows that seeking novelty describes the behavior of human participants better than seeking information-gain and surprise, but it does not tell us which aspects of data statistics cannot be explained by algorithms driven by information-gain or surprise. To investigate this question, we use our three intrinsically motivated algorithms with their fitted parameters and simulate new participants, i.e., we perform Posterior Predictive Checks (PPC)[55,56]. As opposed to the simulations in Fig. 3.4.2, we do not freely choose the level of exploration in simulations for PPC. Rather, the level of exploration of each newly simulated participant is completely determined by the previously fitted parameters from one of the 57 human participants; specifically, each simulated participant belongs to one of the three groups of human participants (e.g., the 3 CHF group), and its action-choices are simulated using a set of parameters fitted to the action-choices of one human participant randomly selected from the participants in that group (Methods).

Given the PPC results, we first perform model-recovery[55] on the data from the simulated participants: Indeed, model recovery confirms that we can infer which algorithm has generated the action-choices of simulated participants (by repeating our model-fitting and -comparison; Fig. 3.4.4B). This implies that even the versions of different algorithms that are closest to human data can be dissociated in our experimental paradigm (average Protected Exceedance Probability[54] ≥ 0.98 for the true model in Fig. 3.4.4B). Next, we perform a systematic comparison between the statistics of the action-choices of human participants and those of the simulated participants (the two most discriminating statistics are reported in Fig. 3.4.5A-B and a systematic analysis in Supplementary Materials). Our results show that simulated participants using novelty as intrinsic rewards reproduce all data statistics (including the zero correlation observed in Fig. 3.4.3A3; see Supplementary Materials), whereas simulated participants using information-gain or surprise fail to do so. The failure of algorithms using information-gain or surprise is most evident regarding the fraction of time spent in the stochastic part during episodes 2-5: 1. We observe that the 2 CHF group of simulated participants who seek information-gain or surprise spends a significantly

smaller fraction of their time (less than half) in the stochastic part of the environment than the 2 CHF group of human participants (Fig. 3.4.5A). 2. Simulated participants using information-gain or surprise fail to reproduce the observed negative correlation between the goal value and the fraction of time spent in the stochastic part (Fig. 3.4.5B). We emphasize that both shortcomings are observed despite the fact that parameters of the algorithms had been previously optimized to explain as best as possible the sequence of action choices across the whole experiment.

The failure of surprise-seeking algorithms to reproduce these statistics is due to the detrimental consequences of seeking surprise in the presence of stochasticity (e.g., as observed for the simulated efficient agents in Episode 5 of Fig. 3.4.2E3). Hence, to stop the simulated participants from spending an enormous amount of time during episode 5 in the stochastic part of the environment, fitting surprise-seeking to action-choices of human participants yields a set of parameters that causes action-choices to be dominated by extrinsic reward (relative importance of surprise-seeking about 0.3 for the 2 CHF group; Fig. 3.4.5C), which in turn cannot explain the overall high level of exploration observed in the 2 CHF group of human participants (Fig. 3.4.5A). Similarly, the relative importance of information-gain is around 0.65 when parameters of a hybrid algorithm driven by information-gain are optimized to fit human behavior. A higher value of relative importance would make, during episode 2, the algorithm too attracted to the stochastic action in state 4 compared to humans (compare Fig. 3.4.2D with Fig. 3.4.3B). With such reduced importance of information-gain, the hybrid algorithm cannot, however, explain the specific behavioral features in Fig. 3.4.5A and B. Therefore, the attraction of human participants to the stochastic part has specific characteristics that are explained by seeking novelty but not by seeking surprise or information-grain.

Taken together our results with simulated participants provide strong quantitative evidence for novelty-seeking as a model of human exploration in our experiment.

Reward optimism correlates with relative importance of novelty

Using novelty-seeking as the most probable model of human behavior, we can now explicitly test our hypothesis that reward optimism increases the relative importance of intrinsic rewards.

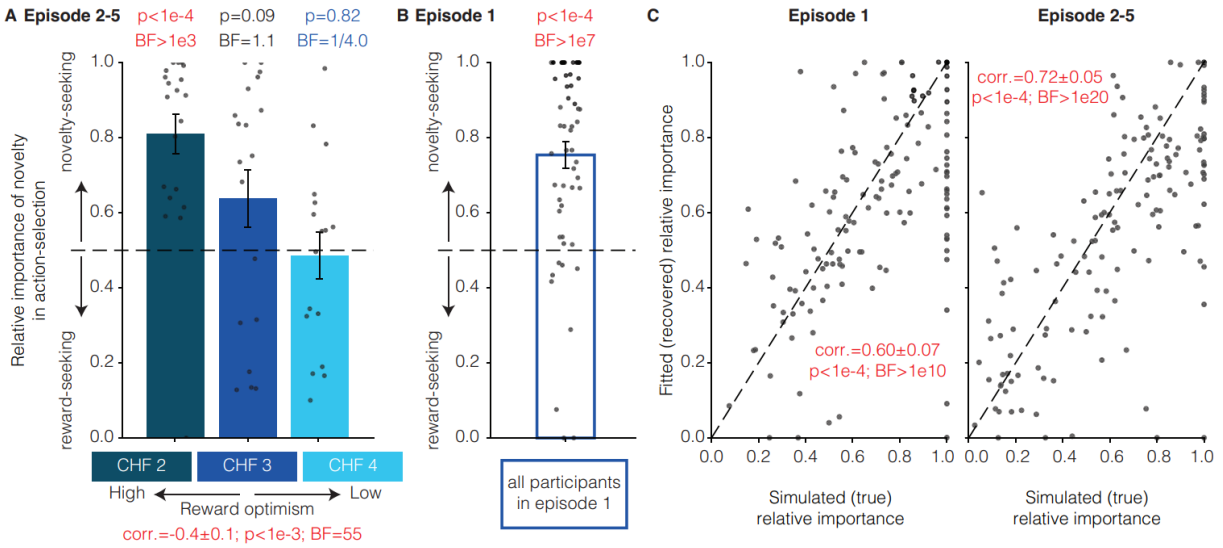

Figure 3.4.6: Reward optimism increases the relative importance of novelty in action-selection. A. The relative importance of novelty-seeking in episodes 2-5 is computed for each participant after fitting the model to data (similar to Fig. **3.4.5C** but using action-choices of human participants instead of simulated participants; Methods). Error bars show the SEMean and single dots the data of individual participants. We observe a significant negative correlation between the relative importance of novelty and the goal value (correlation test; t = −3.6; 95% CI = (−0.63, −0.20); DF = 55). P-values and BFs on top: Comparison with 0.5 (one-sample t-test). We observe a significant dominance of $\pi_{int,t}$ for the 2 CHF group (t = 5.9; 95% CI = (0.70, 0.92); DF = 20). A BF of 1/4.0 in favor of the null hypothesis[49] suggests an equal contribution of $\pi_{ext,t}$ and $\pi_{int,t}$ for the 4 CHF group (t = −0.23; 95% CI = (0.35, 0.62); DF = 16). The test for 3 CHF group is inconclusive (t = 1.8; 95% CI = (0.48, 0.80); DF = 18). **B.** The relative importance of novelty-seeking in episode 1 implies a significant dominance of novelty-seeking against optimistic initialization for exploration (t = 7.3; 95% CI = (0.68, 0.82); DF = 56). **C.** Parameter-recovery[55] using the action-choices of 150 (= 50 per group) simulated participants seeking novelty (Methods). The comparison between the true contribution of novelty-seeking to action-selection (computed with the parameters used for simulations) and the recovered contribution (computed with the parameters fitted to the simulated action-choices) shows that the relative importance of novelty-seeking is on average identifiable in our experimental paradigm: Positive correlations both for episode 1 (t = 9.0; 95% CI = (0.48, 0.70); DF = 148) and episodes 2-5 (t = 12; 95% CI = (0.63, 0.79); DF = 148). Red p-values: Significant effects with False Discovery Rate controlled at 0.05[50] (see Methods). Red BFs: Significant evidence in favor of the alternative hypothesis (BF≥ 3). Blue BFs: Significant evidence in favor of the null hypothesis (BF≤ 1/3).

By analyzing the parameters of our novelty-seeking algorithm fitted to the behavioral data, we observe, in agreement with our hypothesis, a significant negative correlation between the relative importance of novelty during action-selection (in episodes 2-5) and the goal value participants found in episode 1 (Fig. 3.4.6A; parameter-recovery[55] in Fig. 3.4.6C). Moreover, the participants in the 2 CHF group continue with an almost fully exploratory policy in episodes 2-5 indicating that they have only a small bias towards exploiting the small but known reward (Fig. 3.4.6A).

Since our simulated participants are informed that there are three different goal states in the environment, the reward-seeking component $\pi_{ext,t}$ of the action-policy can also contribute to exploratory behavior, e.g., through optimistic initialization of Q-values[37] or prior assumptions about the state-transitions (see Methods for a theoretical analysis). To study the extent of this contribution, we focus on episode 1 where this effect is most easily detectable: We observe a dominant influence of novelty-seeking on action-selection (Fig. 3.4.6B). This implies that, to explain human behavior, the knowledge of the existence of different goal states must drive exploration through a novelty-seeking policy instead of the optimistic initialization of a reward-seeking policy.

3.5 *Discussion*

We designed a novel experimental paradigm to study human curiosity-driven exploration in the presence of stochasticity. We made two main observations: (i) Human participants who are optimistic about finding higher rewards than those already discovered are persistently distracted by stochasticity; and (ii) this persistent pattern of distraction is explained better by seeking novelty than seeking information-gain or surprise, even though seeking information-gain is theoretically more robust in dealing with stochasticity.

How humans deal with the exploration-exploitation trade-off has been a long-lasting question in neuroscience and psychology[57,58]. Experimental studies have shown that humans use a combination of random and directed exploration[13,16], potentially linked to different neural mechanisms[59–61]. However, there are multiple distinct theoretical models to describe directed exploration[5,10,26,62–65], and it has been debated which one is best suited to explain human behavior. In a general setting, human exploration is driven by multiple motivational signals[15,65], but it has been also shown that a particular signal can dominate exploration in specific tasks[3,17,45,66–68]. In an earlier work[18], we have shown that novelty signals dominantly drive human exploration in situations where one needs to search for rewarding states in unknown but deterministic environments. Observations (i) and (ii) above provide further evidence for novelty as the dominant drive of human search strategy even in situations where seeking novelty is not optimal and leads to distraction by reward-independent stochasticity. Further experimental studies are needed to investigate the role novelty in other types of human exploratory behavior.

Observation (ii) is particularly surprising as it has been believed that humans are not prone to the 'noisy TV' problem[4,31,33]. Our results with human participants challenge the idea of defining curiosity as a normative solution to the exploration-exploitation trade-off[3,6]; hence, algorithmic advances in machine learning do not necessarily help finding better models of human exploration. However, we note that, for computing novelty, an agent only needs to track the state frequencies over time and does not need any knowledge of the environment's structure (Methods); hence computing novelty is computationally cheaper than computing information-gain. This suggests that a potentially higher level of distraction by noise in humans may be the price of spending less computational power. In other words, novelty-seeking in the presence of stochasticity may not be a globally optimal strategy for exploration but can be an optimal strategy given a set of prior assumptions and computational constraints, i.e., a 'resource rational' policy[69–71].

The core assumption of using intrinsically motivated algorithms as models of human curiosity is the existence of an intrinsic exploration reward parallel to the extrinsic reward. There are, however, multiple ways to incorporate intrinsic rewards into the RL framework [26]. The common practice is to use a weighted sum of the intrinsic and the extrinsic reward as a single scalar reward signal driving action-selection[8,19,28]. An alternative approach is to treat different reward signals in parallel and compute separate action-policies which are combined to drive action-selection later in the processing stream[18,72,73]. The latter approach provides higher flexibility to arbitrate between different policies based on changes in the relative importance of intrinsic versus extrinsic rewards because it does not need re-planning or re-evaluation of already learned policies. Parallel processing paths are compatible with the rapid change of behavior observed in our human participants after finding a goal at the end of episode 1 (Fig. 3.4.3B1-2) and also consistent with experimental evidence for partially separate neural pathways of novelty- and reward-induced behaviors[18,74–78].

We found that the relative importance of novelty- and reward-induced behaviors in human participants is correlated with the degree of reward optimism. This is in line with the known influence of environmental variables on an agent's preference for novelty[45,46,76]. In particular, theories of 'motivation crowding effect'[79] and 'undermining effect'[80,81] suggest that the absolute value of extrinsic reward might contribute, in addition to the reward optimism, to the observed negative correlation in Fig. 3.4.6A, predicting that even if participants were confident that there is no other goal state in the environment, the 2 CHF group would spend more time in the stochastic part than the 4 CHF group – simply because 2 CHF is not an attractive reward anyway. A potential future direction to investigate the interplay of novelty and reward is to study human behavior in various environments with different reward distributions and different sources of stochasticity.

Optimism in psychology has been defined as a 'variable that reflects the extent to which people hold generalized favorable expectancies for their future'[48] and has been linked to several neural and behavioral characteristics[48,82,83]. While the traditional approach to measure optimism is through self-tests[84], more recently statistical inference using RL[85] and Bayesian[86,87] models of behavior have been proposed to quantify variables correlated with traditional measurements. While there are multiple traditional ways to incorporate the notion of optimism into the RL framework (Methods), seeking intrinsic rewards has also been interpreted in the machine learning community as an 'optimistic policy' for exploration[62]. Our results show that the preference for an

intrinsic reward is indeed correlated with a notion of optimism defined in the context of our experiment as the expectancy of finding a goal of higher value in episodes 2-5 ('reward optimism' in Fig. 3.4.6A). Moreover, the persistent exploration of the stochastic part of our environment observed in the behavior of human participants (Fig. 3.4.3B3) is consistent with the known phenomena of optimism bias[88] and optimistic belief updating in humans[82,89,90].

Even though notions of 'novelty', 'surprise', and 'information-gain' are frequently used in neuroscience[18,91,92], psychology[93,94], and machine learning[26,27,33], there is no consensus on the precise definitions of these notions as scientific terms [44,95]. Our results in this paper are based on the specific mathematical formulations that we have chosen (Methods), but we expect our conclusions to be invariant with respect to the precise choice of definitions as long as (i) novelty quantifies infrequency of states[18], e.g., defined based on density models in machine learning[19,20]; (ii) surprise quantifies mismatches between observations and agents' expectations, where the expectations are made based on the previous state-action pair, including all measure of prediction surprise[44] and typical measures of prediction error in machine learning[21,28]; and (iii) information-gain quantifies improvements in the agents' world-model and vanishes by accumulation of experience, e.g., including Bayesian[91] and Postdictive surprise[92] and measures of disagreement and progress-rate in machine learning[23–25,29].

In conclusion, our results show (i) that human decision-making is influenced by an interplay of intrinsic with extrinsic rewards that is controlled by reward optimism and (ii) that novelty-seeking RL algorithms can successfully model this interplay in tasks where humans search for rewarding states.

Acknowledgement

AM thanks Vasiliki Liakoni, Johanni Brea, Sophia Becker, Martin Barry, Valentin Schmutz, and Guillaume Bellec for many useful discussions on relevant topics. This research was supported by Swiss National Science Foundation No. CRSII2 147636 (Sinergia, MHH and WG), No. 200020 184615 (WG), and No. 200020 207426 (WG) and by the European Union Horizon 2020 Framework Program under grant agreement No. 785907 (Human Brain Project, SGA2, MHH and WG).

Author Contributions

AM, HAX, MHH, and WG developed the study concept and designed the experiment. HAX and WL conducted the experiment and collected the data. AM designed the algorithms, did the formal analyses, and analyzed the data. AM, MHH, and WG wrote the paper.

Competing Interests statement

The authors declare no competing interests.

Code and data availability

All code and data needed to reproduce the results reported in this manuscript will be made publicly available after publication acceptance.

3.6 *References*

1. Berlyne, D. E. A theory of human curiosity. British Journal of Psychology. General Section 45, 180–191 (1954).

2. Berlyne, D. E. Novelty and curiosity as determinants of exploratory behaviour. British Journal of Psychology. General Section 41, 68–80 (1950).

3. Dubey, R. & Griffiths, T. L. Reconciling novelty and complexity through a rational analysis of curiosity. Psychological Review 127, 455–476 (2019).

4. Gottlieb, J., Oudeyer, P.-Y., Lopes, M. & Baranes, A. Information-seeking, curiosity, and attention: computational and neural mechanisms. Trends in Cognitive Sciences 17, 585–593 (2013).

5. Schulz, E. & Gershman, S. J. The algorithmic architecture of exploration in the human brain. Current opinion in neurobiology 55, 7–14 (2019).

6. Singh, S., Lewis, R. L., Barto, A. G. & Sorg, J. Intrinsically motivated reinforcement learning: An evolutionary perspective. IEEE Transactions on Autonomous Mental Development 2, 70–82 (2010).

7. Santucci, V., Baldassarre, G. & Mirolli, M. Which is the best intrinsic motivation signal for learning multiple skills? Frontiers in Neurorobotics 7 (2013).

8. Jaegle, A., Mehrpour, V. & Rust, N. Visual novelty, curiosity, and intrinsic reward in machine learning and the brain. Current Opinion in Neurobiology 58, 167–174 (2019).

9. Oudeyer, P.-Y., Kaplan, F. & Hafner, V. V. Intrinsic motivation systems for autonomous mental development. IEEE Transactions on Evolutionary Computation 11, 265–286 (2007).

10. Murayama, K. A reward-learning framework of knowledge acquisition: An integrated account of curiosity, interest, and intrinsic–extrinsic rewards. Psychological Review 129, 175–198 (2022).

11. Gottlieb, J. & Oudeyer, P.-Y. Towards a neuroscience of active sampling and curiosity. Nature Reviews Neuroscience 19, 758–770 (2018).

12. Kakade, S. & Dayan, P. Dopamine: generalization and bonuses. Neural Networks 15, 549–559 (2002).

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

14. Iigaya, K., Story, G. W., Kurth-Nelson, Z., Dolan, R. J. & Dayan, P. The modulation of savouring by prediction error and its effects on choice. eLife 5, e13747 (2016).

15. Kobayashi, K., Ravaioli, S., Baran`es, A., Woodford, M. & Gottlieb, J. Diverse motives for human curiosity. Nature human behaviour 3, 587–595 (2019).

16. Gershman, S. J. Uncertainty and exploration. Decision 6, 277 (2019).

17. Horvath, L. et al. Human belief state-based exploration and exploitation in an information-selective symmetric reversal bandit task. Computational Brain & Behavior (2021).

18. Xu, H. A., Modirshanechi, A., Lehmann, M. P., Gerstner, W. & Herzog, M. H. Novelty is not surprise: Human exploratory and adaptive behavior in sequential decision-making.PLoS Computational Biology 17 (2021).

19. Bellemare, M. et al. Unifying count-based exploration and intrinsic motivation. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I. & Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 29 (Curran Associates, Inc., 2016).

20. Ostrovski, G., Bellemare, M. G., van den Oord, A. & Munos, R. Count-based exploration with neural density models. In Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, 2721–2730 (JMLR.org, 2017).

21. Pathak, D., Agrawal, P., Efros, A. A. & Darrell, T. Curiosity-driven exploration by self-supervised prediction. In Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, 2778–2787 (JMLR.org, 2017).

22. Haber, N., Mrowca, D., Wang, S., Fei-Fei, L. F. & Yamins, D. L. Learning to play with intrinsically-motivated, self-aware agents. In Bengio, S. et al. (eds.) Advances in Neural Information Processing Systems, vol. 31 (Curran Associates, Inc., 2018).

23. Sekar, R. et al. Planning to explore via self-supervised world models. In III, H. D. & Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning, vol. 119 of Proceedings of Machine Learning Research, 8583–8592 (PMLR, 2020).

24. Kim, K., Sano, M., De Freitas, J., Haber, N. & Yamins, D. Active world model learning with progress curiosity. In III, H. D. & Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning, vol. 119 of Proceedings of Machine Learning Research, 5306–5315 (PMLR, 2020).

25. Mendonca, R., Rybkin, O., Daniilidis, K., Hafner, D. & Pathak, D. Discovering and achieving goals via world models. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. & Vaughan, J. W. (eds.) Advances in Neural Information Processing Systems, vol. 34, 24379–24391 (Curran Associates, Inc., 2021).

26. Aubret, A., Matignon, L. & Hassas, S. A survey on intrinsic motivation in reinforcement learning. arXiv preprint arXiv:1908.06976 (2019).

27. Ladosz, P., Weng, L., Kim, M. & Oh, H. Exploration in deep reinforcement learning: A survey. Information Fusion 85, 1–22 (2022).

28. Burda, Y. et al. Large-scale study of curiosity-driven learning. In International Conference on Learning Representations (2019).

29. Pathak, D., Gandhi, D. & Gupta, A. Self-supervised exploration via disagreement. In Chaudhuri, K. & Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning, vol. 97 of Proceedings of Machine Learning Research, 5062–5071 (PMLR, 2019).

30. Savinov, N. et al. Episodic curiosity through reachability. In International Conference on Learning Representations (2019).

31. Mavor-Parker, A., Young, K., Barry, C. & Griffin, L. How to stay curious while avoiding noisy TVs using aleatoric uncertainty estimation. In Chaudhuri, K. et al. (eds.) Proceedings of the 39th International Conference on Machine Learning, vol. 162 of Proceedings of Machine Learning Research, 15220–15240 (PMLR, 2022).

32. Jarrett, D. et al. Curiosity in hindsight. In Deep Reinforcement Learning Workshop NeurIPS 2022 (2022).

33. Schmidhuber, J. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). IEEE Transactions on Autonomous Mental Development 2, 230–247 (2010).

34. Tartaglia, E. M., Clarke, A. M. & Herzog, M. H. What to choose next? a paradigm for testing human sequential decision making. Frontiers in Psychology 8, 312 (2017).

35. Lehmann, M. P. et al. One-shot learning and behavioral eligibility traces in sequential decision making. eLife 8, e47463 (2019).

36. Liakoni, V. et al. Brain signals of a surprise-actor-critic model: Evidence for multiple learning modules in human decision making. NeuroImage 246, 118780 (2022).

37. Sutton, R. S. & Barto, A. G. Reinforcement learning: An introduction (MIT press, 2018).

38. Daw, N., Gershman, S., Seymour, B., Dayan, P. & Dolan, R. Model-based influences on humans' choices and striatal prediction errors. Neuron 69, 1204–1215 (2011).

39. Huys, Q. J. et al. Interplay of approximate planning strategies. Proceedings of the National Academy of Sciences 112, 3098–3103 (2015).

40. Van Seijen, H. & Sutton, R. Planning by prioritized sweeping with small backups. In Dasgupta, S. & McAllester, D. (eds.) Proceedings of the 30th International Conference on Machine Learning, vol. 28 of Proceedings of Machine Learning Research, 361–369 (PMLR, Atlanta, Georgia, USA, 2013).

41. Mattar, M. G. & Lengyel, M. Planning in the brain. Neuron 110, 914–934 (2022).

42. Mobin, S. A., Arnemann, J. A. & Sommer, F. Information-based learning by agents in unbounded state spaces. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. & Weinberger, K. Q. (eds.) Advances in Neural Information Processing Systems, vol. 27 (Curran Associates, Inc., 2014).

43. Little, D. Y.-J. & Sommer, F. T. Learning and exploration in action-perception loops. Frontiers in Neural Circuits 7, 37 (2013).

44. Modirshanechi, A., Brea, J. & Gerstner, W. A taxonomy of surprise definitions. Journal of Mathematical Psychology 110, 102712 (2022).

45. Gershman, S. J. & Niv, Y. Novelty and inductive generalization in human reinforcement learning. Topics in cognitive science 7, 391–415 (2015).

46. Stoji´c, H., Schulz, E., P Analytis, P. & Speekenbrink, M. It's new, but is it good? how generalization and uncertainty guide the exploration of novel options. Journal of Experimental Psychology: General 149, 1878–1907 (2020).

47. Traner, M. R., Bromberg-Martin, E. S. & Monosov, I. E. How the value of the environment controls persistence in visual search. PLoS Computational Biology 17, 1–34 (2021).

48. Carver, C. S., Scheier, M. F. & Segerstrom, S. C. Optimism. Clinical Psychology Review 30, 879–889 (2010). Positive Clinical Psychology.

49. Kass, R. E. & Raftery, A. E. Bayes factors. Journal of the american statistical association 90, 773–795 (1995).

50. Efron, B. & Hastie, T. Computer age statistical inference (Cambridge University Press, 2016).

51. Daw, N., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nature neuroscience 8, 1704–1711 (2005).

52. Gl¨ascher, J., Daw, N., Dayan, P. & O'Doherty, J. P. States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron 66, 585–595 (2010).

53. Daw, N. Trial-by-trial data analysis using computational models. Decision making, affect, and learning: Attention and performance XXIII 23 (2011).

54. Rigoux, L., Stephan, K. E., Friston, K. J. & Daunizeau, J. Bayesian model selection for group studies—revisited. Neuroimage 84, 971–985 (2014).

55. Wilson, R. C. & Collins, A. G. Ten simple rules for the computational modeling of behavioral data. eLife 8, e49547 (2019).

56. Nassar, M. R. & Frank, M. J. Taming the beast: extracting generalizable knowledge from computational models of cognition. Current opinion in behavioral sciences 11, 49–54 (2016).

57. Cohen, J. D., McClure, S. M. & Yu, A. J. 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, 933–942 (2007).

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

59. Zajkowski, W. K., Kossut, M. & Wilson, R. C. A causal role for right frontopolar cortex in directed, but not random, exploration. eLife 6, e27430 (2017).

60. Dubois, M. et al. Human complex exploration strategies are enriched by noradrenaline-modulated heuristics. eLife 10, e59907 (2021).

61. Wittmann, B. C., Daw, N. D., Seymour, B. & Dolan, R. J. Striatal activity underlies novelty-based choice in humans. Neuron 58, 967–973 (2008).

62. Ghavamzadeh, M., Mannor, S., Pineau, J. & Tamar, A. Bayesian reinforcement learning: A survey. Found. Trends Mach. Learn. 8, 359–483 (2015).

63. Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P. & Pezzulo, G. Active inference: a process theory. Neural computation 29, 1–49 (2017).

64. Klyubin, A., Polani, D. & Nehaniv, C. Empowerment: a universal agent-centric measure of control. In 2005 IEEE Congress on Evolutionary Computation, vol. 1, 128–135 Vol.1 (2005).

65. Br¨andle, F., Stocks, L. J., Tenenbaum, J. B., Gershman, S. J. & Schulz, E. Intrinsically motivated exploration as empowerment. PsyArXiv (2022).

66. Ten, A., Kaushik, P., Oudeyer, P.-Y. & Gottlieb, J. Humans monitor learning progress in curiosity-driven exploration. Nature communications 12, 5972 (2021).

67. Itti, L. & Baldi, P. Bayesian surprise attracts human attention. Vision Research 49, 1295–1306 (2009).

68. Meder, B. & Nelson, J. D. Information search with situation-specific reward functions. Judgment and Decision Making 7, 119–148 (2012).

69. Lieder, F. & Griffiths, T. L. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. Behavioral and Brain Sciences 43, e1 (2020).

70. Bhui, R., Lai, L. & Gershman, S. J. Resource-rational decision making. Current Opinion in Behavioral Sciences 41, 15–21 (2021). Value based decision-making.

71. Binz, M. & Schulz, E. Modeling human exploration through resource-rational reinforcement learning. In Oh, A. H., Agarwal, A., Belgrave, D. & Cho, K. (eds.) Advances in Neural Information Processing Systems (2022).

72. Kim, Y., Nam, W., Kim, H., Kim, J.-H. & Kim, G. Curiosity-bottleneck: Exploration by distilling task-specific novelty. In Chaudhuri, K. & Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning, vol. 97 of Proceedings of Machine Learning Research, 3379–3388 (PMLR, 2019).

73. Cogliati Dezza, I., Cleeremans, A. & Alexander, W. H. Independent and interacting value systems for reward and information in the human brain. eLife 11, e66358 (2022).

74. Ghazizadeh, A. et al. Brain Networks Sensitive to Object Novelty, Value, and Their Combination. Cerebral Cortex Communications 1 (2020). Tgaa034.

75. Menegas, W., Babayan, B. M., Uchida, N. & Watabe-Uchida, M. Opposite initialization to novel cues in dopamine signaling in ventral and posterior striatum in mice. eLife 6, e21886 (2017).

76. Akiti, K. et al. Striatal dopamine explains novelty-induced behavioral dynamics and individual variability in threat prediction. Neuron 110, 3789–3804.e9 (2022).

77. Ogasawara, T. et al. A primate temporal cortex–zona incerta pathway for novelty seeking. Nature neuroscience 25 (2022).

78. Tapper, A. R. & Molas, S. Midbrain circuits of novelty processing. Neurobiology of Learning and Memory 176, 107323 (2020).

79. Frey, B. S. & Jegen, R. Motivation crowding theory. Journal of Economic Surveys 15, 589–611 970 (2001).

80. Deci, E. L., Koestner, R. & Ryan, R. M. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychological Bulletin 125, 627–668 (1999).

81. Murayama, K., Matsumoto, M., Izuma, K. & Matsumoto, K. Neural basis of the undermining effect 974 of monetary reward on intrinsic motivation. Proceedings of the National Academy of Sciences 107, 975 20911–20916 (2010).

82. Sharot, T., Korn, C. W. & Dolan, R. J. How unrealistic optimism is maintained in the face of reality. 977 Nature neuroscience 14, 1475–1479 (2011).

83. Strunk, D. R., Lopez, H. & DeRubeis, R. J. Depressive symptoms are associated with unrealistic negative predictions of future life events. Behaviour Research and Therapy 44, 861–882 (2006).

84. Scheier, M. F., Carver, C. S. & Bridges, M. W. Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A reevaluation of the life orientation test. Journal of Personality and Social Psychology 67, 1063–1078 (1994)

85. Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S. & Palminteri, S. Behavioural and neural characterization of optimistic reinforcement learning. Nature Human Behaviour 1, 1–9 (2017).

86. Stankevicius, A., Huys, Q. J. M., Kalra, A. & Seri`es, P. Optimism as a prior belief about the probability of future reward. PLoS Computational Biology 10, 1–9 (2014).

87. Gesiarz, F., Cahill, D. & Sharot, T. Evidence accumulation is biased by motivation: A computational account. PLoS Computational Biology 15, 1–15 (2019).

88. Sharot, T. The optimism bias. Current Biology 21, R941–R945 (2011).

89. Garrett, N. & Sharot, T. Optimistic update bias holds firm: Three tests of robustness following shah et al. Consciousness and Cognition 50, 12–22 (2017). Unrealistic Optimism – Its Nature, Causes and Effects.

90. Palminteri, S. & Lebreton, M. The computational roots of positivity and confirmation biases in 994 reinforcement learning. Trends in Cognitive Sciences 26, 607–621 (2022).

91. Baldi, P. A Computational Theory of Surprise, 1–25 (Springer US, Boston, MA, 2002).

92. Kolossa, A., Kopp, B. & Fingscheidt, T. A computational analysis of the neural bases of bayesian inference. NeuroImage 106, 222–237 (2015).

93. Reisenzein, R., Horstmann, G. & Sch¨utzwohl, A. The cognitive-evolutionary model of surprise: A review of the evidence. Topics in Cognitive Science 11, 50–74 (2019).

94. Nelson, J. D. Finding useful questions: on bayesian diagnosticity, probability, impact, and information gain. Psychological Review 112, 979–999 (2005).

95. Barto, A., Mirolli, M. & Baldassarre, G. Novelty or surprise? Frontiers in Psychology 4, 907 (2013).

96. Brainard, D. H. & Vision, S. The psychophysics toolbox. Spatial vision 10, 433–436 (1997).

97. Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D. & Iverson, G. Bayesian t tests for accepting and rejecting the null hypothesis. Psychonomic bulletin & review 16, 225–237 (2009).

98. Rouder, J. N. & Morey, R. D. Default bayes factors for model selection in regression. Multivariate Behavioral Research 47, 877–903 (2012).

99. Ghahramani, Z. Bayesian non-parametrics and the probabilistic approach to modelling. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 371, 20110553 (2013).

100. Yu, A. J. & Cohen, J. D. Sequential effects: Superstition or rational behavior? In Koller, D., Schuurmans, D., Bengio, Y. & Bottou, L. (eds.) Advances in Neural Information Processing Systems, vol. 21 (Curran Associates, Inc., 2009).

101. Liakoni, V., Modirshanechi, A., Gerstner, W. & Brea, J. Learning in volatile environments with the bayes factor surprise. Neural Computation 33, 1–72 (2021).

102. Piray, P. & Daw, N. D. Linear reinforcement learning in planning, grid fields, and cognitive control. Nature communications 12, 4942 (2021).

103. Cover, T. M. Elements of information theory (John Wiley & Sons, 1999).

104. Rowan, T. H. Functional stability analysis of numerical algorithms. Ph.D. thesis, The University of Texas at Austin (1990).

105. Johnson, S. G. The nlopt nonlinear-optimization package. URL http://github.com/stevengj/nlopt

4. Study II: Abnormal dopamine levels lead to deteriorated reward but enhanced punishment learning in schizophrenia patients

In our first study, we aimed to examine how abnormalities in the dopamine system affect an individual's ability in RL. Previous research has reported abnormally elevated dopamine levels in patients with schizophrenia. However, there has been no direct explanation of how such deficits affect RL. We propose that elevated dopamine levels restrict the dopamine transient, leading to diminished neural activity for encoding the prediction error. This decrease in the prediction error signal could affect the reward update process, resulting in suboptimal action selection in RL. Additionally, our model suggests that elevated dopamine levels might not impact punishment learning, as the negative prediction error is unaffected by elevated baseline dopamine levels. Therefore, our hypotheses are: 1) Schizophrenia patients will exhibit deficits in reward-based learning, and 2) these patients will demonstrate intact punishment-based learning.

To validate these hypotheses, we designed an RL task in which participants were presented with an image (state) and four circles (action) below it. Participants had to click one circle to transition to another state, until a predefined goal state was reached. Our results indicated that the patients were less accurate compared to the healthy controls, and the cause is primarily due to increased perseverative behavior. This pattern was also revealed in our computer simulations. Second, when punishment is introduced into the RL environment, the patients performed better compared to the reward condition, supporting our second hypothesis.

Abnormal dopamine levels lead to deteriorated reward but enhanced punishment learning in schizophrenia patients

Wei-Hsiang Lin[1], Maya Roinishvili[2,3], Mariam Okruashvili[6], Eka Chkonia[4,5] & Michael H. Herzog[1]

1. Laboratory of Psychophysics, Brain Mind Institute, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
2. Laboratory of Vision Physiology, Ivane Beritashvili Centre of Experimental Biomedicine, Tbilisi, Georgia
3. School of Science and Technology, University of Georgia, Tbilisi, Georgia
4. Department of Psychiatry, Tbilisi State Medical University (TSMU) Tbilisi, Georgia
5. Institute of Cognitive Neurosciences, Free University of Tbilisi, Tbilisi, Georgia
6. Mental Health Center, Tbilisi, Georgia

4.1 *Abstract*

Dopamine (DA) is a critical neurotransmitter in reinforcement learning (RL) processes (Schultz et al., 1997), and its aberrantly high levels in striatal region are pivotal in schizophrenia. However, the link between the learning deficits observed in schizophrenia and dopamine dysregulation has remained a puzzle. In this study, we propose a novel hypothesis that compressed transient dopamine signals, resulting from the elevated dopamine levels in schizophrenia patients, disrupt learning. Our theoretical model generates three predictions, all of which are supported by our results. First, we observe that patients exhibit suboptimal performance in RL tasks. Second, our computational simulations suggest that the compression of transient dopamine signals induces perseverative behavior in patients, thereby explaining their impaired task performance. Third, both our theoretical predictions and empirical data indicate that schizophrenia patients have preserved, or even enhanced, RL capabilities in tasks that involve punishment rather than reward.

4.2 *Introduction*

Dopamine is essential to reinforcement learning (RL). Notably, dopaminergic neurons in the pars compacta of the substantia nigra exhibit robust firing in response to unexpected rewards and diminished firing for anticipated ones, regardless of the reward's magnitude (Schultz et al., 1997). Such neural activities illustrate the concept of reward prediction error, a fundamental component of all RL models. This prediction error is typically characterized as the difference between the anticipated and received rewards. A plethora of research supports the notion that transient fluctuations in dopamine concentrations significantly influence synaptic neural plasticity (Calabresi et al., 2007; Centonze et al., 2001; Shen et al., 2008; van der Schaaf et al., 2014).

Dysfunctions in the dopamine system are frequently implicated as a critical factor in psychiatric disorders, and more specifically in schizophrenia (Kapur et al., 2005; Howes and Kapur, 2009a). For instance, D2 receptor antagonists have been found to alleviate psychotic symptoms (Abi-Dargham et al., 1998; Davidson et al., 1987), whereas dopamine agonists can induce psychotic symptoms even in healthy individuals (Janowsky and Risch, 1979; Sekine et al., 2001). Research by Seeman (2013) suggests that psychosis may arise when D2 receptors adopt a high-affinity state, pointing to these receptors as potential mediators of learning effects. This aligns with the hypothesis that excessive dopamine uptake by neurons, facilitated by these D2 receptors, correlates with the issues related to dopamine transients that we've highlighted. As a result, many therapeutic approaches for schizophrenia predominantly focus on modulating dopamine D2 receptors to stabilize aberrant dopamine levels (Howes et al., 2012; Howes and Kapur, 2009b). Nonetheless, the precise effects of dopamine in schizophrenia are multifaceted and remain inadequately understood.

Elevated baseline striatal dopamine levels have been consistently observed in patients with schizophrenia (Kesby et al., 2018; Seeman, 2011; Davis et al., 1991; Laruelle and Abi-Dargham, 1999). We proposed that these anomalously high levels lead to a compressed encoding of the prediction error (Fig. 4.4.1A), resulting in the learning deficits observed in these patients. Specifically, this elevated baseline may diminish the contrast between baseline and peak dopamine levels, thus reducing the precision of prediction error encoding and subsequently slowing the RL process in patients. To explore this hypothesis, we have modified the standard Q-learning model, a method wherein an agent iteratively navigates an environment, taking actions that transition from one state to another and updating values based on prediction errors following each action. We

introduced a term to represent the effect of compressed dopamine transients on prediction error:

$$Q(S,A) \leftarrow Q(S,A) + \alpha\left(\beta\big(r - Q(S,A)\big) + \max_i Q(S_{new}, A_i)\right)$$

In the equation, $Q(S,A)$ represents the value updated for each state (S) and action (A). The parameter β serves as a scaling factor to modulate the prediction error ($r - Q(S,A)$), where r is the reward. A smaller β implies greater compression of the prediction error signal, reflecting the hypothesized effect of abnormal dopamine transients. α is the learning rate. The latter part of the update rule, $\max_i Q(S_{new}, A_i)$, specifies the maximum expected Q-value in the next state (S_{new}), over all possible actions (A_i). This component serves as an estimate of future value, guiding the agent towards actions that are expected to yield the highest rewards in subsequent states.

To validate our model and formulate predictions, we incorporated it into a behavioral RL paradigm. In this paradigm, participants are presented with an image (state) accompanied by four gray buttons (Fig 4.4.1; Tartaglia et al., 2017). Participants select a button (action), prompting the appearance of a subsequent image. This process is repeated until a target image appears. The sequence and nature of these images, representing the "environment", resists simplification into a straightforward 2D map. For instance, pressing the right button might transition from a cherry image to a house image, but no single action brings the display back to the cherry. We specifically designed this paradigm to prevent participants from forming cognitive maps, thus ensuring our setup closely mirrors fundamental RL scenarios while minimizing cognitive demands.

In our simulation, we adjusted β to either 75% or 50% of its maximum value and ran the simulation in the same environment as that used for human participants. Our analyses considered the average number of repeated state-action pairs, the maximum number of repetitions within an episode, and the length of the repeating patterns. Our findings suggest that compressed prediction errors may lead to increased perseveration behavior (Fig. 4.4.2), a characteristic commonly observed in schizophrenia patients (Crider, 1997; Ersche et al. 2011). Additionally, our model suggests that patients might perform better in RL with negative rewards – particularly punishments – compared to positive ones. This is due to a potentially larger prediction error signal with negative rewards, which we expressed as "max dopamine level - β*baseline dopamine level". Since we presume β for patients is less than one, this calculation yields a larger discrepancy than in controls. This implies that patients could potentially outperform controls in scenarios driven by punishment.

To validate these hypotheses, we conducted two experiments using the aforementioned paradigm, with 20 schizophrenia patients and 20 controls, respectively.

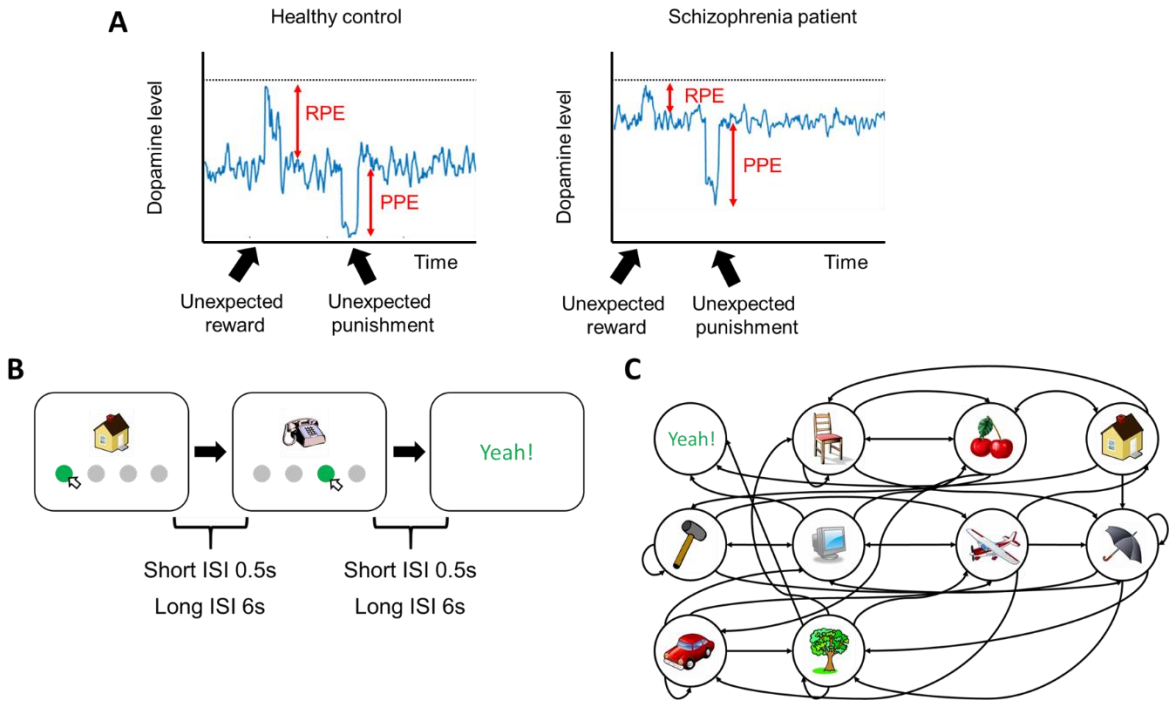

Figure 4.4.1. The RL task. (A): Model overview. The depicted model showcases differential baseline dopamine levels in healthy controls (left) versus schizophrenia patients (right). Due to elevated baseline levels in patients, the range of dopamine transients that encode the reward prediction error (RPE) is comparatively compressed. Conversely, when considering punishments, the dynamics are inverted. (**B)**: Illustration of a single episode in the experimental environment. Participants encountered an initial image, for instance, a house, accompanied by four gray buttons. Upon selecting a button (an action), they transitioned to a subsequent image, like the telephone. This progression continued until participants reached the "Yeah!" image. The succession of images and corresponding button presses leading to the "Yeah!" image is termed an episode. Participants endeavored to finish as many episodes as possible within a designated timeframe. Image inter-stimulus intervals (ISI) vary between 0.5s and 6s depending on the experimental conditions. (**C)**: State-action connectivity representation. Each node symbolizes an image, with arrows indicating potential actions. The environment structure is notably irregular. As an illustration, while there's a path from the house image to the cherries, no direct return link exists. Mathematically, this environment cannot be simplified into a 2D map, a design intended to suppress cognitive factors (Tartaglia et al., 2017).

4.3 *Methods*

Participants

We recruited 20 schizophrenia patients and 20 age-matched healthy controls from the Asatiani Psychiatric Institute in Tbilisi, Georgia, for experiments 1 and 2, respectively. Patient diagnoses adhered to DSM-IV, determined through a combination of SCID-based interviews, staff input, and a review of patient records. Psychopathology was assessed by the Scales for the Assessment of Negative and Positive Symptoms (SANS, SAPS respectively; Andreasen, 1989, 1984). Demographic details are presented in Table 1. All participating patients were under neuroleptic treatment. Chlorpromazine equivalents (CPZ) are shown in Table 1.

	Experiment 1		Experiment 2	
	Patients	Controls	Patients	Controls
Gender (F/M)	1/19	10/10	3/17	5/15
Age	40.6 ± 8.4	29.9 ± 7.3	43.9 ± 8	44 ± 6.1
Education	14.1 ± 3.2	16.6 ± 3.8	12.6 ± 2.6	14.3 ± 3.4
Illness duration	17.9 ± 8	--	20.5 ± 8.9	--
SANS	11.1 ± 4.6	--	10.8 ± 5.1	--
SAPS	8.8 ± 2.7	--	8.8 ± 3.4	--
CPZ	525.3±416.3	--	559.5 ± 328.6	--

Table 1. Average Statistics (±SD) of Schizophrenic Patients and Healthy Controls. Abbreviations: SANS, Scale for the Assessment of Negative Symptoms; SAPS, Scale for the Assessment of Positive Symptoms, CPZ, Chlorpromazine equivalents.

Procedure: Experiment 1

Each participant was first presented with a screen containing all nine images used in the experiment, with a gray button positioned at the bottom. This button served as the start button for the experiment. Once initiated, the participant was presented with one image centered on the screen and with four gray buttons below. Selecting one of these buttons brought the participant to another image until the goal image, labeled with the word “Yeah!”, was found (Fig. 1, top).

We refer to the sequence of images (states) and mouse clicks (actions) leading up to and including the “Yeah!” (goal) as one episode. We used two inter-stimulus-intervals (ISIs) of 0.5 seconds and 6 seconds between the observers’ response and the next image presentation. The task was to find the goal state as often as possible in 8 minutes for the 0.5-second ISI condition, and 30 minutes for the 6-second ISI condition.

Procedure: Experiment 2

Experiment 2 was similar to experiment 1 but incorporated both reward and punishment conditions, with an ISI of 0.5 seconds. In the reward condition, participants aimed to reach the goal state 15 times, with each successful attempt earning a monetary reward of 1 CHF, for a potential total of 15 CHF. Conversely, the punishment condition imposed monetary penalties for each fourth "incorrect" action within six pre-specific states. Participants were alerted with the message "Wrong way" and incurred a loss of 1 CHF for every fourth occurrence of entering a punishment state. This condition was also completed after 15 episodes. Notably, these punishment states were positioned at least two steps away from the goal state. The order of the two conditions was pseudo-randomized among participants, ensuring that half began with the reward condition and the other half with the punishment condition.

Analysis

Number of Episodes and proportion of optimal actions

Performance was quantified by the number of episodes each participant completed within the given timeframe, with a higher episode count indicating better performance. Since observers varied in reaction times, leading to differences in episode completion, we standardized this metric. We identified the minimum number of total actions performed by any participant in each condition. Subsequently, for each participant, we considered only this minimum number of actions, disregarding any actions beyond this group minimum. One patient was excluded from the analysis due to completing fewer than 1st percentile of the number of actions relative to other participants.

Additionally, we evaluated performance based on the proportion of optimal actions taken by each participant. This was calculated by dividing the number optimal actions by the total number of actions. An action was deemed optimal if it minimized the distance from the current state to the goal state.

Perseveration

Schizophrenia patients are known to exhibit perseverative behavior, characterized by their tendency to engage in repetitive, non-optimal action sequences. For instance, they might repeatedly perform actions such as "1, 2, 3, 1, 2, 3, 4, 2 …". To quantify this behavior, we computed the frequency and the length of repeating segments within each action sequence for

every participant in each episode. A repeating segment is defined as a sequence of actions that occurs more than once (e.g., "1, 2, 3" in the above sequence, which has a length of three). In instances of multiple repeating segments within an episode, our analysis focused on the segment that was repeated most frequently (Fig. S1).

Q-learning

In our experimental setup, we defined an environment comprising s = 10 states, each with a = 4 possible actions. To calculate the probability q of selecting each of the four possible actions (indexed by i), we utilized the softmax selection rule:

$$q(i) = \frac{exp\left(\frac{Q(S,A_i)}{\tau}\right)}{\sum_j exp\left(\frac{Q(S,A_j)}{\tau}\right)}$$

Here, q(i) specifies the probability of selecting action Ai when in state S, while τ denotes the exploration rate (higher τ values signify increased exploration). The Q-matrix is updated as follows:

$$Q(S,A) \leftarrow Q(S,A) + \alpha\left(r + \max_i Q(S_{new},A_i) - Q(S,A)\right)$$

The methodology allowed us to extract the maximum likelihood estimates for the parameters α (learning rate) and τ (exploration rate) from each participant's data.

Hypothesis and simulation

To model the compressed dopamine transient characteristic of schizophrenia, we employed a Q-learning algorithm across 100 simulations, introducing a specific modification to the Q-update rule as follows:

$$Q(S,A) \leftarrow Q(S,A) + \alpha\left(\beta\left(r - Q(S,A)\right) + \max_i Q(S_{new},A_i)\right)$$

Here, β represents a scalar that varies from zero to one, indicating the proportion of the dopaminergic error signal (r-Q(S,A)) that is accessible for reinforcement learning. A β value of one signifies full availability of the dopaminergic error signal, while zero implies complete unavailability. Contrary to constraining the model to follow participants' actions, we allowed it to autonomously select actions based on the aforementioned soft-max selection rule. This was simulated by setting the scaling factor, β, to either 0.75 or 0.5, corresponding to a 25% or 50%

reduction in the maximum error signal. We then assessed the maximum perseveration values and analyzed the distribution of perseveration values across different β settings. This analysis was aimed at understanding the impact of the compressed maximum error signal on perseveration behavior (Fig. 4.4.2).

4.4 *Results*

Simulations

Our simulation's findings underscore that compressing the prediction error augments perseverative behavior. The model unveiled a higher proportion of repeated state-action pairs when error signals were compressed (Fig. 4.4.2A; left: $F(2,297) = 45.05$, $p < 0.001$, partial $\eta 2 = 0.23$, post-hoc analysis: $t_{0\% - 25\%} = -6.11$, $p < 0.001$, $t_{25\% - 50\%} = -3.23$, $p = 0.004$; Fig. 4.4.2B; left: $F(2,297) = 98.49$, $p < 0.001$, partial $\eta 2 = 0.4$, post-hoc analysis: $t_{0\% - 25\%} = -9.91$, $p < 0.001$, $t_{25\% - 50\%} = -3.65$, $p < 0.001$). Additionally, the distribution for both the proportion of repeated state-action pairs and repeated lengths was more prominent when the prediction error remained uncompressed, in comparison to when compressed to 75% or 50% (as per the β parameter in our Q-learning adjustment). The Kolmogorov-Smirnov (KS) test corroborated these findings (Fig 4.4.2A; right: $KS_{0\% \text{ vs. } 25\%} = 0.57$, $p < 0.001$, $KS_{25\% \text{ vs. } 50\%} = 0.47$, $p < 0.001$, $KS_{0\% \text{ vs. } 50\%} = 0.85$, $p < 0.001$; Fig 4.4.2B; right: $KS_{0\% \text{ vs. } 25\%} = 0.94$, $p < 0.05$, $KS_{25\% \text{ vs. } 50\%} = 0.48$, $p < 0.001$, $KS_{0\% \text{ vs. } 50\%} = 0.97$, $p < 0.001$).

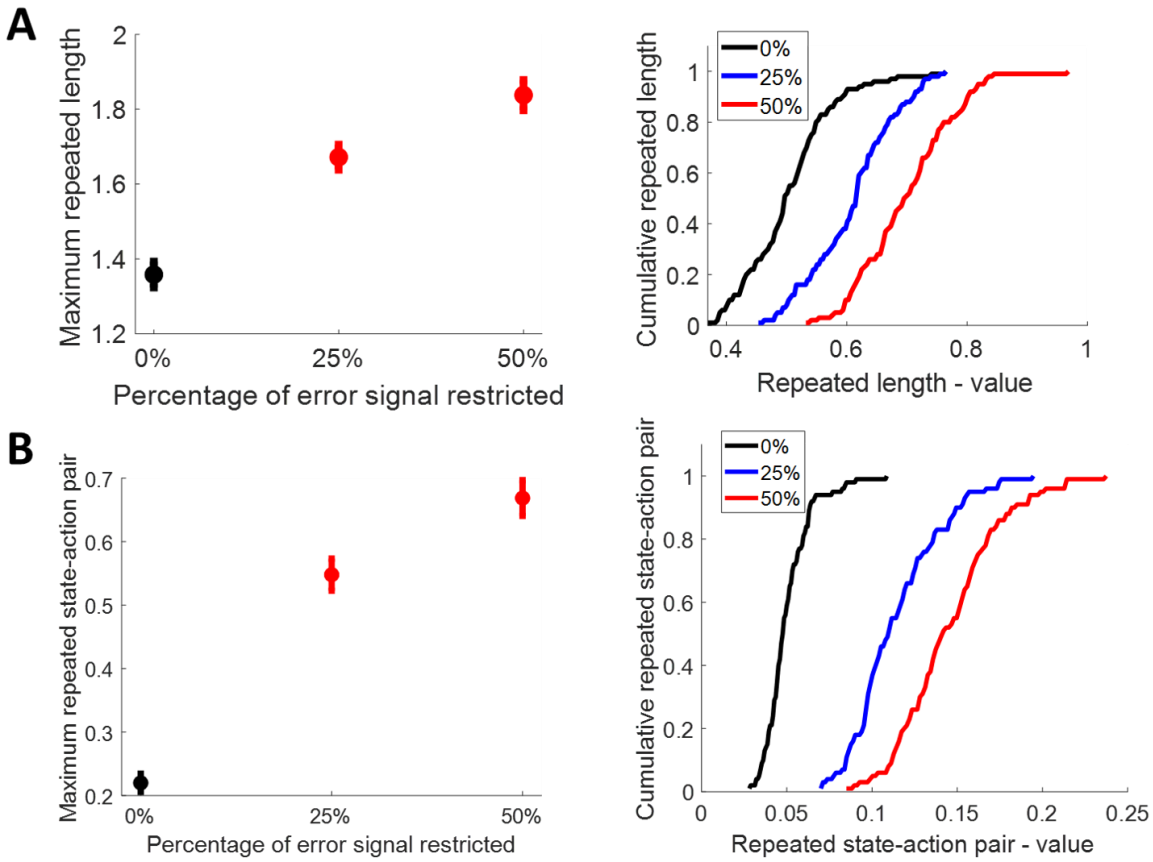

Figure 4.4.2. Simulations. (A): Repeated length. Left: The maximum repeated length as a function of the percentage of error signal compressed. Right. The cumulative probability density of the distribution of the simulated repeated length. (B): Proportion of repeated state-action pairs. Left: The maximum proportion of repeated state-action pair as a function of the percentage of error signal compressed. Right. The cumulative probability density of the distribution of the simulated proportion of repeated state-action pair. Error bars represent ±1 SEM.

Experiment 1

Effects of ISI. A 2 × 2 ANOVA (Illness: healthy, schizophrenia; ISI: 0.5s, 6s) revealed a significant main effect of illness ($F(1,37) = 29.94$, $p < 0.001$, partial $\eta 2 = 0.45$) and ISI ($F(1,37) = 6.45$, $p = 0.02$, partial $\eta 2 = 0.15$; Fig. 4.4.3A). The interaction between illness and ISI was marginally significant ($F(1,37) = 3.87$, $p = 0.06$, partial $\eta 2 = 0.01$). Fig. 4.4.3B shows that patients took fewer optimal actions than healthy controls ($F(1,37) = 32.78$, $p < 0.001$, partial $\eta 2 = 0.47$) and a main effect of ISI ($F(1,37) = 6.2$, $p = 0.02$, partial $\eta 2 = 0.14$).

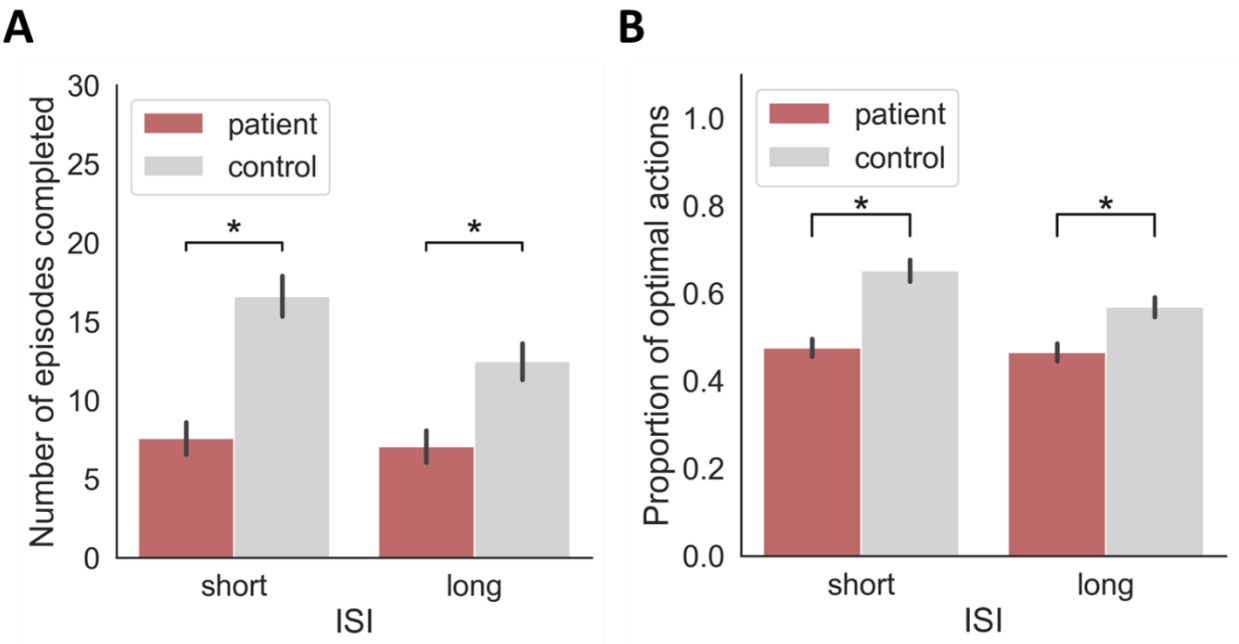

Figure 4.4.3. Performance in the RL task. (A): Number of completed episodes as a function of ISI for patients with schizophrenia and healthy controls. There is a significant difference in the number of completed episodes between the two groups at both ISI conditions. **(B)**: Proportion of optimal action as a function of ISI for patients with schizophrenia and healthy controls. There is a significant difference in the proportion of optimal actions between the two groups at both ISI conditions. Error bars represent ±1 SEM. Stars indicate the main effect of groups (patient vs. control, $p < 0.05$).

Exploration and Perseverations. Exploration was quantified as the mean number of actions taken at each state. Given that there are four potential actions per state, this average can range from a maximum of four (if all actions were chosen at least one) to a minimum of zero (for states that were never visited). Patients with schizophrenia exhibited greater exploration compared to the controls (Fig. 4.4.4A; $F(1,37) = 4.89$, $p = 0.03$, partial $\eta 2 = 0.12$). As the ISI lengthened, exploration also increased ($F(1,37) = 12.62$, $p = 0.001$, partial $\eta 2 = 0.25$). A marginal interaction was observed between the illness and ISI ($F(1,37) = 3.97$, $p = 0.054$, partial $\eta 2 = 0.1$).

While patients with schizophrenia explored more than controls, their exploration was often inefficient, frequently neglecting optimal actions (Fig. 4.4.3B). Possible explanations for this include a tendency to perseverate or potential memory impairments. To assess the degree of participants' perseverative behavior, we evaluated the average number of times each participant revisited the same state-action pair within an episode. Revisiting a particular state within the same episode inherently indicates a suboptimal strategy, as optimal task navigation eliminates the need

for such repetition. Fig. 4.4.4B revealed that the probability of patients with schizophrenia repeating state-action pairs was not statistically different from that of healthy controls ($F(1,37) = 2.62$, $p = 0.11$, partial $\eta 2 = 0.07$). However, the patient group demonstrated a longer sequence of repeated actions ($F(1,37) = 6.66$, $p = 0.01$, partial $\eta 2 = 0.15$) than controls (Fig. 4.4.4C). These results suggest that patients with schizophrenia not only engage in heightened exploration but also exhibit more pronounced perseveration tendencies compared to healthy controls.

While we cannot entirely rule out the potential influence of memory deficits on participants' ability to recall previously chosen actions upon revisiting a state, we believe such deficits have a minimal impact. This assertion is based on our observation that performance in conditions with a long ISI was not significantly different from that in conditions with a short ISI.

The pronounced tendency of patients to perseverate could account for their sub-optimal performance. We also observed that patients often shift between repetitive patterns across episodes, which accounts for their elevated exploration rates. They initiate a sequence of state-action choices, but instead of progressing towards the goal, they become trapped in recurrent cycles, struggling to escape. In subsequent episodes, they initiate new cycles that incorporate different states, thereby expanding the variety of state-action pairs they engage with. However, these new patterns do not lead to improved performance, as they too fail to effectively guide patients towards the goal state.

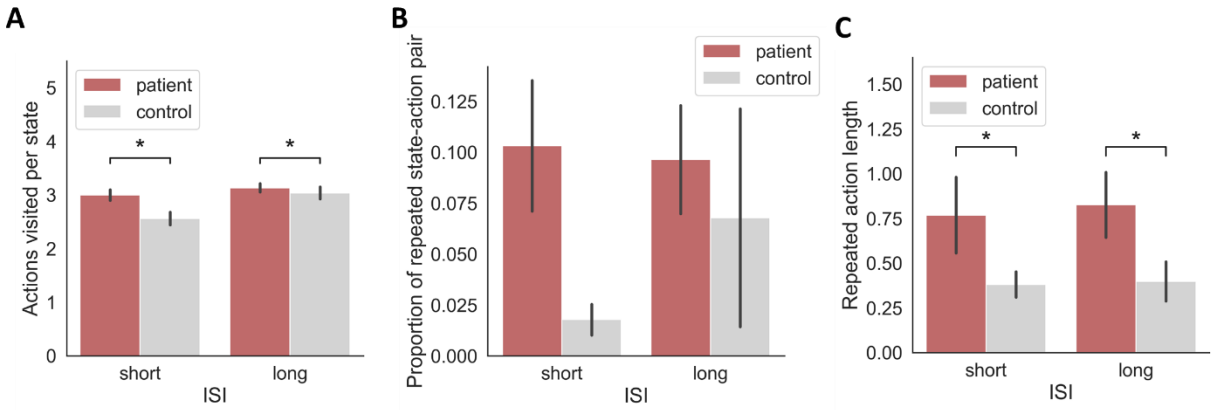

Figure 4.4.4. Exploration and perseveration. (A): Proportion of actions visited per state at both ISI conditions. A main effect of illness was observed, indicating that the patients explored more actions compared to controls. (B): Mean number of repeated cycles per episode at both ISI conditions. No significant difference was observed between the two groups. (C): Mean length of repeating segment averaged over episodes at both ISI conditions. A significant difference was observed in the repeated action length between the two groups for both ISIs. Error bars represent ±1 SEM. Stars indicate the main effect of groups (patient vs. control, $p < 0.05$).

Q-learning. A strength of our experimental paradigm is its compatibility with reinforcement learning models, which enable us to dissect the impact of various factors on performance. We derived parameter estimates by fitting the trial-by-trial decisions of each participant across episodes, yielding exceptionally robust estimates. The reinforcement learning model employed in this study, Q-learning, has two parameters: learning rate (alpha) and exploratory rate (tau). Our findings reveal no significant main effect of illness on the learning rate (Fig. 4.4.5A; $F(1,37) = 0.26$, $p = 0.61$, partial $\eta 2 = 0.007$). However, there is a significant effect of illness on the exploration rate (Fig. 4.4.5B; $F(1,37) = 26.47$, $p < 0.001$, partial $\eta 2 = 0.42$), with patients exhibiting higher exploration rates compared to healthy controls.

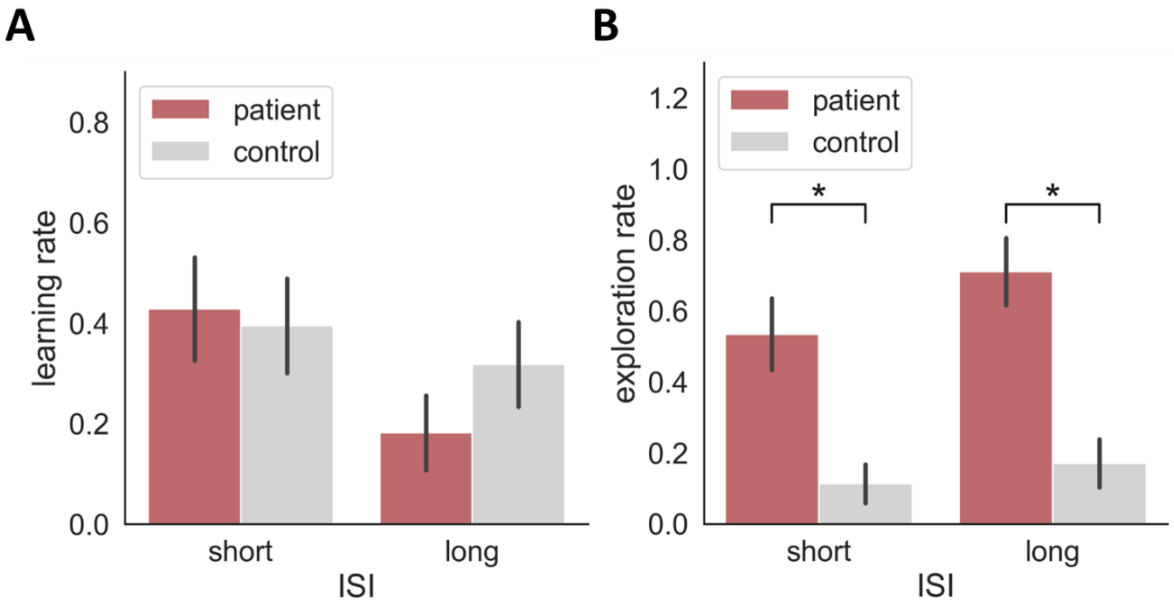

Figure 4.4.5. Q-learning parameters. (A): Learning rates. (B): Exploration rates. The plots show that patients with schizophrenia have very similar learning rates as control, but exhibit more exploratory behavior. Error bars represent ±1 SEM. Stars indicate the main effect of groups (patient vs. control, $p < 0.05$).

Experiment 2

Effects of punishment. We conducted a second experiment, building upon our hypothesis that patients demonstrate superior performance under punishment conditions compared to rewards.

This is attributed to elevated baseline levels of dopamine, which we propose leads to more effective encoding of negative prediction errors (Fig. 4.4.1A). The aim of this study was to assess how reward and punishment differentially impact RL performance in patients.

First, we observed significant differences between conditions concerning both the number of episodes completed (Fig. 4.4.6A; $t(19) = -2.74$, $p = 0.01$, Cohen's $d = 0.61$) and the proportion of optimal actions (Fig. 4.4.6B; $t(19) = 2.71$, $p = 0.01$, Cohen's $d = 0.61$). These findings suggest that overall accuracy was higher in conditions involving punishment compared to those with rewards. Additionally, we investigated the influence of these conditions on perseveration behavior. Interestingly, no significant differences were found in the levels of perseveration, indicating that both reward and punishment may impact perseveration similarly (Fig. 4.4.6C; $t(19) = 1.74$, $p = 0.1$, Cohen's $d = 0.39$). However, the data indicate a trend towards reduced perseverative behaviors in the punishment condition as compared to the reward condition among patients.

Patients with schizophrenia exhibited distinct responses to the reward and punishment conditions (Fig. 4.4.6), a difference that might be attributed to the experimental design itself. To determine if this effect was specific to the patient population, we conducted tests with an age-matched healthy control group, using identical procedures and measurements. Interestingly, in the control group, we observed no significant differences in either accuracy (Fig. 4.4.S2A; $t(19) = -0.97$, $p = 0.35$, Cohen's $d = 0.22$) or perseveration behavior (Fig. 4.4.S2B; $t(19) = 0.74$, $p = 0.46$, Cohen's $d = 0.17$) across the two conditions. These results suggest that the two experimental conditions were indeed comparable for the control group.

Notably, the performance of the healthy control group varied between experiment 1 and 2. This discrepancy may be due to the inherent nature of the tasks, since we implemented preprocessing to equalize the number of actions performed by participants within each experiment. With 71 actions for experiment 1 and 53 actions for experiment 2, a direct comparison between these experiments is not feasible. Furthermore, as the hypothesis and objectives for each experiment were distinct, it would be inappropriate to directly compare the two experiments, considering they served different purposes.

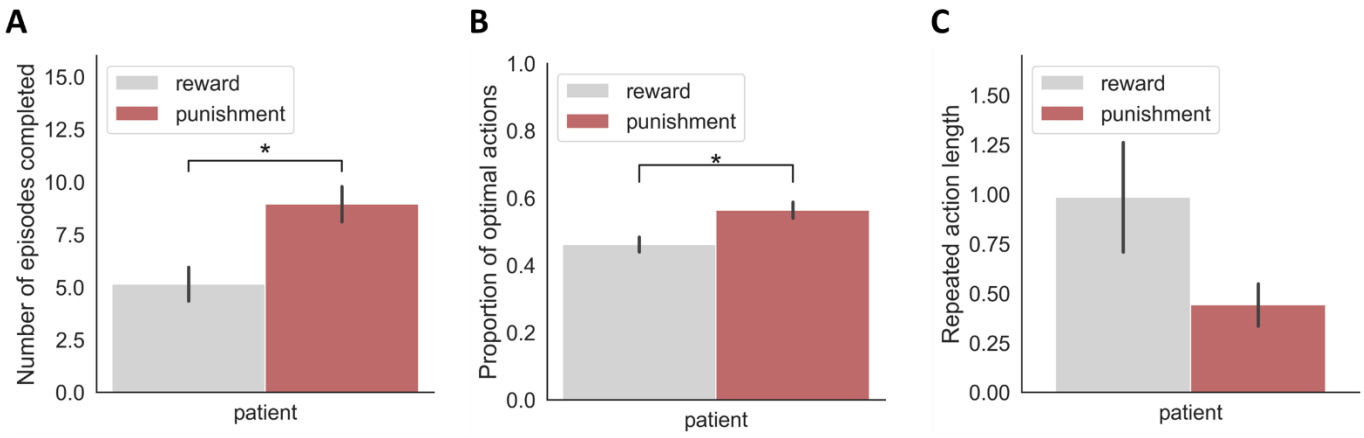

Figure 4.4.6. Performance in the RL task. (A) The number of episodes completed by patients with schizophrenia in reward and punishment conditions. Patients completed more episodes in the punishment condition compared to the reward condition. **(B):** The proportion of optimal actions of patients with schizophrenia in reward and punishment conditions. Patients had a higher proportion of optimal actions in the punishment condition compared to the reward condition. **(C)**: The length of repeated actions of patients with schizophrenia in reward and punishment conditions. There is no difference in the perseveration behavior of patients between the reward and punishment conditions. Error bars represent ±1 SEM. Stars indicate a significant difference between the two experimental conditions ($p < 0.05$).

Q-learning. In line with the first experiment, we utilized the Q-learning model to quantify the performance of the patients. First, we did not observe significant differences in learning rates between the reward and punishment conditions, suggesting comparable learning across conditions (Fig. 4.4.7A; $t(19) = -0.35$, $p = 0.73$, Cohen's $d = 0.08$). Furthermore, no differences were observed in the exploration rate between the two conditions (Fig. 4.4.7B; $t(19) = 1.17$, $p = 0.26$, Cohen's $d = 0.26$), reinforcing the idea that patients' action selection remains consistent with the RL task, irrespective of the experimental conditions.

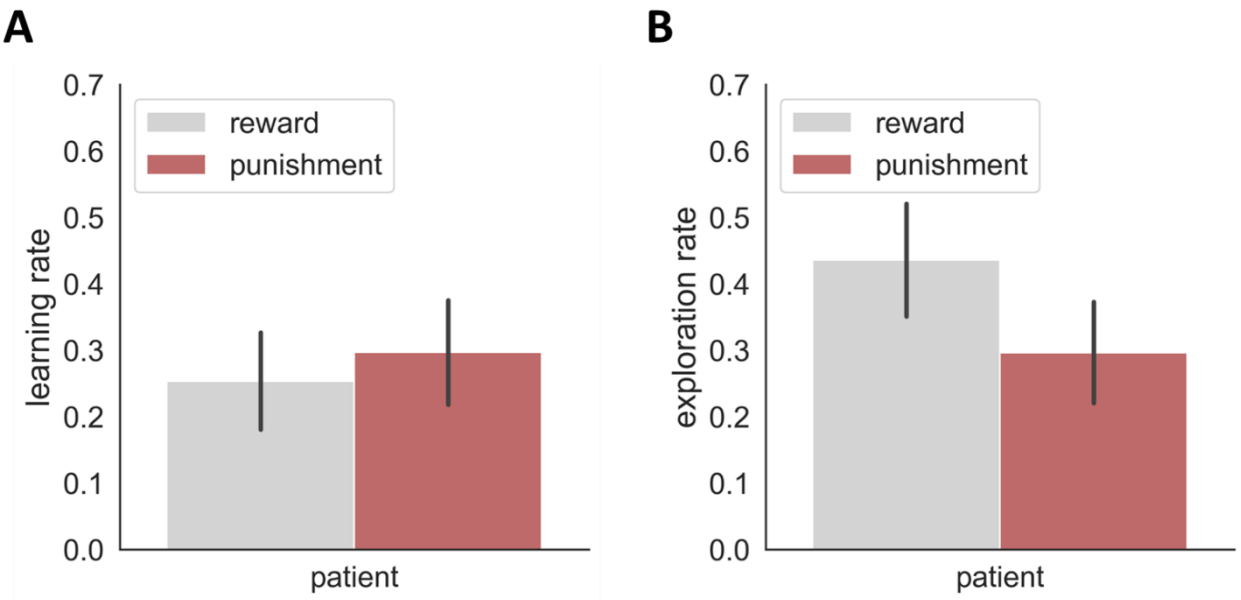

Figure 4.4.7. Q-learning parameters. (A): Learning rates. (B): Exploration rate. Patients with schizophrenia have very similar learning rates and exploration rates in the two experimental conditions. Error bars represent ±1 SEM.

4.5 *Discussion*

Elevated striatal dopamine levels are widely believed to play a pivotal role in schizophrenia, especially given that psychotic symptoms can be alleviated by dopamine antagonists and most medications target the D2 receptor. However, the precise mechanism linking elevated dopamine levels to symptoms remains unclear. Dopamine is known to encode the prediction error, i.e., the difference between the actual and expected reward, which is, for this reason, the crucial component in most RL models. Individuals with schizophrenia show significant impairments in RL. Here, we suggested that a modified RL model can account for these deficits. The key idea is that elevated baseline levels of dopamine lead to less efficient encoding of the prediction error due to the reduced range between the baseline and the maximal level compared to controls. Our model makes three predictions: 1) patients with schizophrenia perform sub-optimally in RL compared to healthy controls with positive reward, 2) this poor performance emerges from action choice, and 3) patients exhibit minimal or no deficits in RL with negative rewards, i.e., punishments. We confirmed these predictions.

Perseveration, which is characterized by the repetition of specific action sequences regardless of their outcomes, is a typical behavioral pattern in patients with schizophrenia and was observed in both experiments conducted here. In the Q-learning model, such perseveration is attributed to a diminished encoding of prediction errors, leading to reduced magnitude of updating values associated with state-action pairs and consequently slower learning. In the model, the inverse temperature parameter (exploration rate) is higher in patients compared to controls. As a result, unlike controls, patients tend to wander through the environment until they happen to hit the goal by chance. We suggest that the model effectively replicates the underlying mechanisms in these patients.

An alternative explanation for the observed perseverations is memory deficits, i.e., patients do not remember well previous actions. In our RL task, we had two different ISIs of 0.5s and 6s. Should memory deficits be a primary issue, we would expect performance to deteriorate in the long ISI condition relative to the short ISI condition. However, there were no major differences.

In the results of experiment 2, patients had significantly higher accuracy in the punishment condition compared to the reward condition. Our results are in line with previous studies, which found impairments in RL with positive reward but not with punishment (Abohamza et al., 2020; Cheng et al., 2012; Waltz et al., 2007). Another study found deficits in forming new stimulus-

reward associations, yet the patients performed normally when forming new stimulus-punishment associations (Waltz et al., 2007).

Maia and Frank (2017) proposed that spontaneous transient increase of dopamine lead to false associations between stimuli and outcomes (Pankow et al., 2015; Roiser et al., 2013), potentially correlating with positive symptoms. Conversely, diminished transient responses to relevant stimuli (Deserno et al., 2013; Winton-Brown et al., 2014) might contribute to negative symptoms.

Here, we demonstrated that elevated baseline levels of dopamine lead to impoverished reinforcement learning in patients with schizophrenia. As a speculation, we propose that imprecise encoding of predictions could also explain a variety of deficits observed in schizophrenia. Future work needs to investigate how these learning deficits can be linked to psychotic symptoms.

Acknowledgements

This project was funded by the Swiss National Science Foundation (SNF) NCCR project Synapsy: The Synaptic Bases of Mental Diseases and the Sinergia project “Learning from Delayed and Sparse Feedback”. We would like to thank Aaron Clarke for his significant contributions to both the model and the experimental design.

Supplementary figure

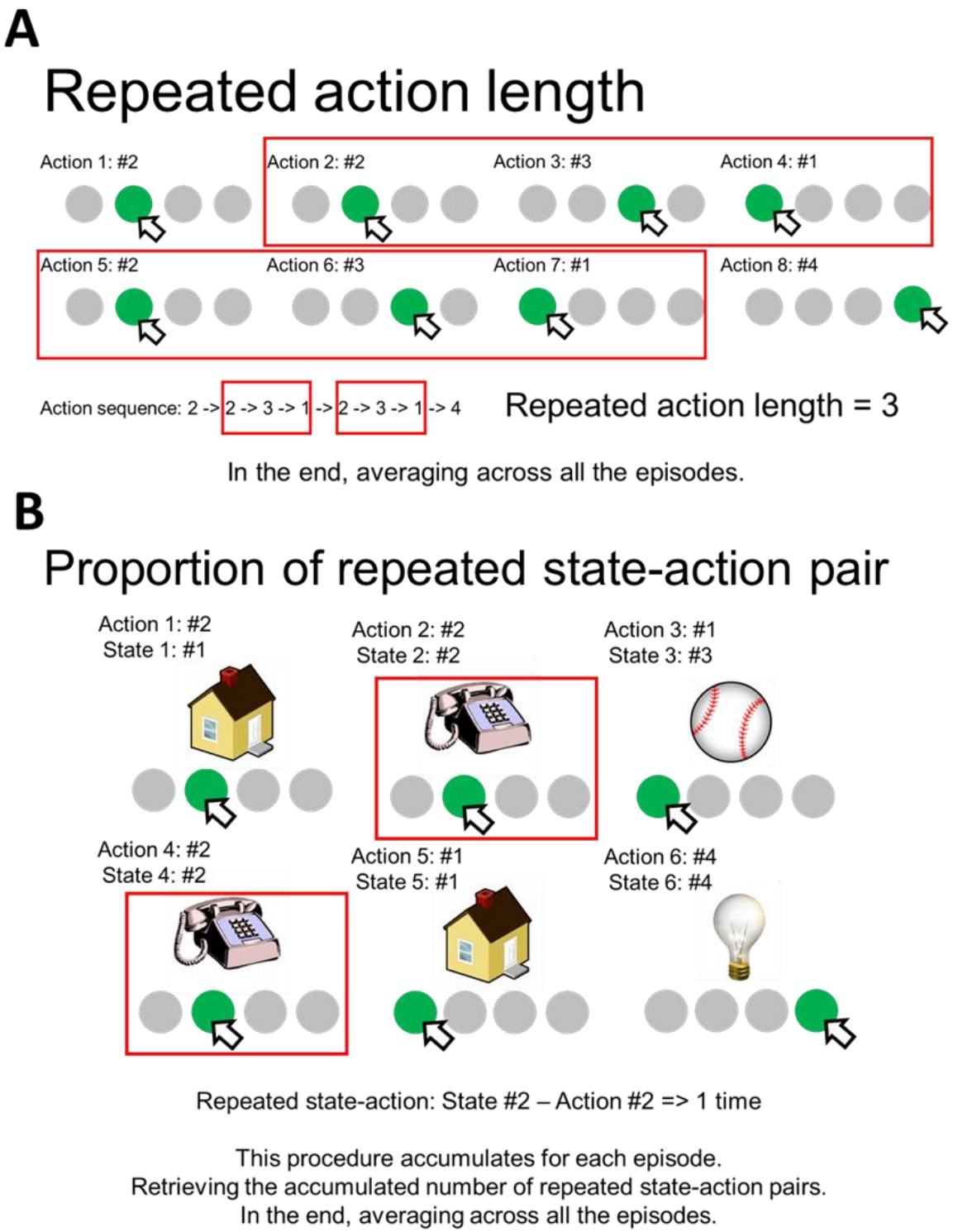

Figure 4.4.S1. Calculation of two measurements of perseveration. (A) The length of repeated actions. **(B)** The proportion of repeated state-action pairs.

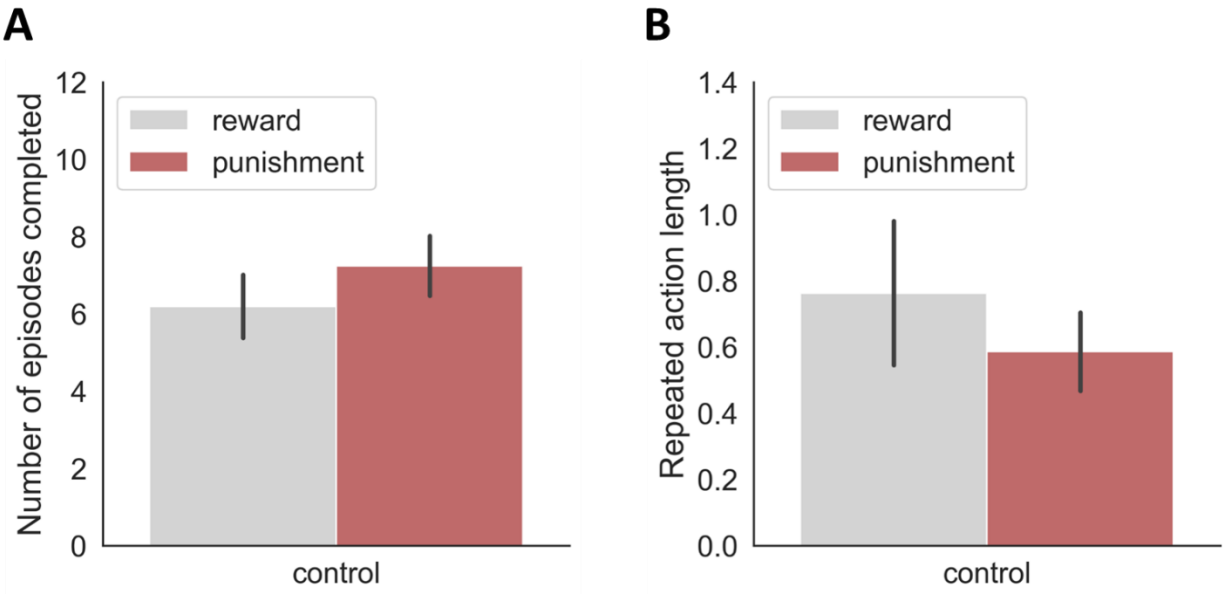

Figure 4.4.S2. Performance on the RL task – Healthy controls. (A) The number of episodes completed by healthy controls in reward and punishment conditions. **(B)**: The repeated length of healthy controls in reward and punishment conditions. Error bars represent ±1 SEM.

4.6 *References*

Abi-Dargham A, Gil R, Krystal J, Baldwin RM, Seibyl JP, Bowers M, van Dyck CH, Charney DS, Innis RB, Laruelle M. 1998. Increased striatal dopamine transmission in schizophrenia: confirmation in a second cohort. Am J Psychiatry 155:761–767.

Abohamza E, Weickert T, Ali M, Moustafa AA. 2020. Reward and punishment learning in schizophrenia and bipolar disorder. Behav Brain Res 381:112298. doi:10.1016/j.bbr.2019.112298

Andreasen NC. 1989. The Scale for the Assessment of Negative Symptoms (SANS): conceptual and theoretical foundations. Br J Psychiatry 155:49–52.

Andreasen NC. 1984. Scale for the assessment of positive symptoms. Psychiatr Psychobiol.

Calabresi P, Picconi B, Tozzi A, Di Filippo M. 2007. Dopamine-mediated regulation of corticostriatal synaptic plasticity. Trends Neurosci 30:211–219.

Centonze D, Picconi B, Gubellini P, Bernardi G, Calabresi P. 2001. Dopaminergic control of synaptic plasticity in the dorsal striatum. Eur J Neurosci 13:1071–1077.

Cheng GLF, Tang JCY, Li FWS, Lau EYY, Lee TMC. 2012. Schizophrenia and risk-taking: Impaired reward but preserved punishment processing. Schizophr Res 136:122–127. doi:10.1016/j.schres.2012.01.002

Crider A. 1997. Perseveration in Schizophrenia. Schizophr Bull 23:63–74. doi:10.1093/schbul/23.1.63

Davidson M, Keefe RS, Mohs RC, Siever LJ, Losonczy MF, Horvath TB, Davis KL. 1987. L-dopa challenge and relapse in schizophrenia. Am J Psychiatry 144:934–938.

Deserno L, Boehme R, Heinz A, Schlagenhauf F. 2013. Reinforcement Learning and Dopamine in Schizophrenia: Dimensions of Symptoms or Specific Features of a Disease Group? Front Psychiatry 4. doi:10.3389/fpsyt.2013.00172

Howes OD, Kambeitz J, Kim E, Stahl D, Slifstein M, Abi-Dargham A, Kapur S. 2012. The Nature of Dopamine Dysfunction in Schizophrenia and What This Means for Treatment: Meta-analysis of Imaging Studies. Arch Gen Psychiatry 69. doi:10.1001/archgenpsychiatry.2012.169

Howes OD, Kapur S. 2009a. The Dopamine Hypothesis of Schizophrenia: Version III--The Final Common Pathway. Schizophr Bull 35:549–562. doi:10.1093/schbul/sbp006

Howes OD, Kapur S. 2009b. The Dopamine Hypothesis of Schizophrenia: Version III--The Final Common Pathway. Schizophr Bull 35:549–562. doi:10.1093/schbul/sbp006

Janowsky DS, Risch C. 1979. Amphetamine psychosis and psychotic symptoms. Psychopharmacology (Berl) 65:73–77.

Kapur S, Mizrahi R, Li M. 2005. From dopamine to salience to psychosis—linking biology, pharmacology and phenomenology of psychosis. Schizophr Res 79:59–68. doi:10.1016/j.schres.2005.01.003

Maia TV, Frank MJ. 2017. An Integrative Perspective on the Role of Dopamine in Schizophrenia. Biol Psychiatry 81:52–66. doi:10.1016/j.biopsych.2016.05.021

Pankow A, Katthagen T, Diner S, Deserno L, Boehme R, Kathmann N, Gleich T, Gaebler M, Walter H, Heinz A, Schlagenhauf F. 2015. Aberrant Salience Is Related to Dysfunctional Self-Referential Processing in Psychosis. Schizophr Bull sbv098. doi:10.1093/schbul/sbv098

Roiser JP, Howes OD, Chaddock CA, Joyce EM, McGuire P. 2013. Neural and Behavioral Correlates of Aberrant Salience in Individuals at Risk for Psychosis. Schizophr Bull 39:1328–1336. doi:10.1093/schbul/sbs147

Schultz W, Dayan P, Montague PR. 1997. A Neural Substrate of Prediction and Reward. Science 275:1593–1599. doi:10.1126/science.275.5306.1593

Seeman P. 2013. Are dopamine D2 receptors out of control in psychosis? Prog Neuropsychopharmacol Biol Psychiatry 46:146–152.

Sekine Y, Iyo M, Ouchi Y, Matsunaga T, Tsukada H, Okada H, Yoshikawa E, Futatsubashi M, Takei N, Mori N. 2001. Methamphetamine-related psychiatric symptoms and reduced brain dopamine transporters studied with PET. Am J Psychiatry 158:1206–1214.

Shen W, Flajolet M, Greengard P, Surmeier DJ. 2008. Dichotomous dopaminergic control of striatal synaptic plasticity. Science 321:848–851.

Tartaglia EM, Clarke AM, Herzog MH. 2017. What to Choose Next? A Paradigm for Testing Human Sequential Decision Making. Front Psychol 8. doi:10.3389/fpsyg.2017.00312

van der Schaaf ME, van Schouwenburg MR, Geurts DE, Schellekens AF, Buitelaar JK, Verkes RJ, Cools R. 2014. Establishing the dopamine dependency of human striatal signals during reward and punishment reversal learning. Cereb Cortex 24:633–642.

Waltz JA, Frank MJ, Robinson BM, Gold JM. 2007. Selective reinforcement learning deficits in schizophrenia support predictions from computational models of striatal-cortical dysfunction. Biol Psychiatry 62:756–764.

Winton-Brown TT, Fusar-Poli P, Ungless MA, Howes OD. 2014. Dopaminergic basis of salience dysregulation in psychosis. Trends Neurosci 37:85–94. doi:10.1016/j.tins.2013.11.003

5. Study III: Intact reinforcement learning in healthy ageing

Daniel et al. (2020) demonstrated that older adults performed less accurately than the younger individuals when tasks were complex, attributing this discrepancy to the increased attentional load affecting the older adults. However, this claim is largely based on a ceiling effect and a non-significant interaction effect, which do not eliminate the possibility that the simple task is not a suitable control in this context. Moreover, we believe there is a universal issue in RL tasks, which are typically too simple, often comprising two images, each representing different probabilities of receiving a reward, thereby allowing for only two states and two actions. Consequently, some studies might not find performance differences between older and young adults simply due to the simplistic nature of the task. To bridge the gap between these varying results, we employed an RL task in Study II that featured a more realistic design compared to those used in other studies. By using this task, we were able to test our primary hypothesis of whether older adults performed less optimally compared to young adults.

In the first experiment in the Study III, we applied the same RL task as Study II to both older and young adults. We here majorly focused on the manipulation of the length of ISI to vary the memory load, with a longer inter-stimulus interval (ISI) requiring a higher memory load in order to retain the stimulus in the memory system. We extracted four behavioral parameters and two parameters (learning and exploration rates) from a classic Q-learning model. We merely observed one of the measurements demonstrating difference between the two age groups (Fig. 5.1A). This gives us confidence to assert that older adults can perform as well as the young population in RL, even when the environment is considerably more complex.

Nevertheless, there might be a possibility that we did not observe a sufficient number of actions to draw a conclusive result. Hence, in the second experiment, we increased the number of actions that participants had to perform in the task. We did observe a difference emerge, but only in two specific behavioral parameters, suggesting that the performance of young adults improved and outperformed the older ones, potentially due to a greater improvement in selecting the optimal actions among the young adults (Fig. 5.1B).

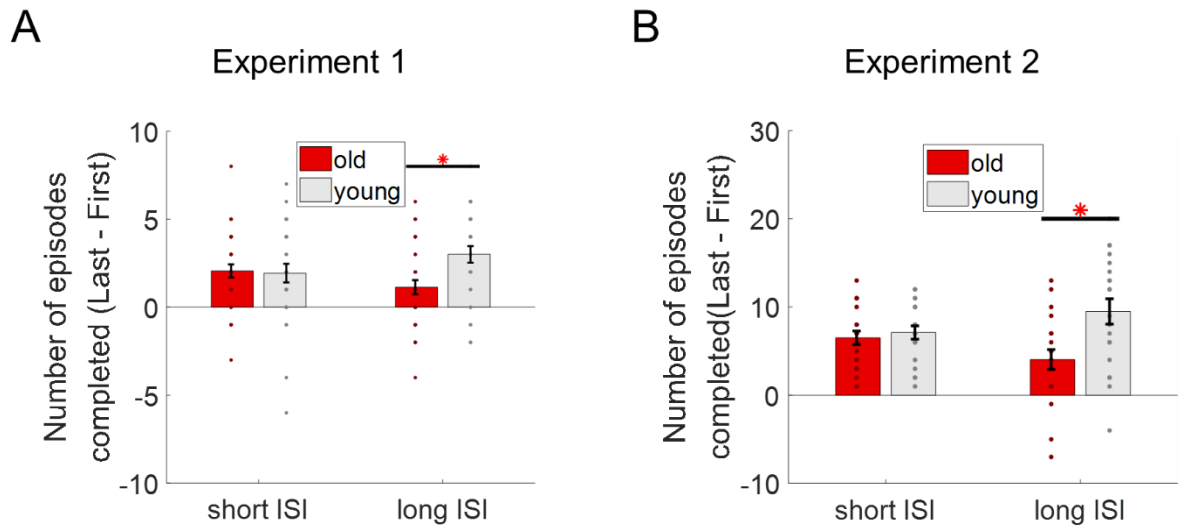

Figure 5.1. Improvement in performance across age groups in two ISI conditions. (A) Measurements from experiment 1 in Study III, with the two color bars indicating the two different age groups. The y-axis represents the difference between the number of episodes completed in the last and first halves of the experiment. (B) The same measurements applied to experiment 2. * indicates $p < 0.05$.

Taking these results together, we conclude that older adults generally exhibit intact RL capabilities comparable to young adults. We believe these findings offer several important perspectives in studies related to aging. Firstly, it is imperative to employ more realistic tasks to accurately find out the true effects. Secondly, and most importantly, we emphasize the criticality of publishing null results in research to avert potential biases that may suggest older adults invariably exhibit certain cognitive deficits.

Intact reinforcement learning in healthy ageing

Wei-Hsiang Lin[1], Karin S. Pilz[2], Michael H. Herzog[1] & Marina Kunchulia[3]

1. Laboratory of Psychophysics, Brain Mind Institute, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
2. Cito Institute for Educational Measurement, Arnhem, The Netherlands
3. Institute of Cognitive Neurosciences, Free University of Tbilisi, Tbilisi, Georgia

5.1 *Abstract*

What does age in ageing? Results in reinforcement learning (RL) are mixed. Some studies found deteriorated performance in older participants compared to younger controls whereas other studies did not. Daniel et al. (2020) suggested that task demand can explain these differences, with less demanding tasks showing no effect of age. Here, we increased the task demand of previous studies turning them into tasks that resemble navigation. We extracted 4 behavioral parameters and 2 parameters (learning and exploration rates) of a classic Q-learning model. Except for one specific parameter, all other parameters showed no group differences, i.e., RL turned out to be intact in older individuals also with higher task demands. It is important to publish such null results to avoid the stigmatizing impression of an overall performance deficit among older people.

5.2 *Introduction*

To understand the mechanisms of ageing, it is of great importance to understand first what functions decline and which do not. Learning is often thought to significantly decay with age (Anguera et al., 2011; Salthouse, 2009). However, evidence is mixed. In reinforcement learning (RL), for example, a clear decline was found by Daniel et al. (2020) and van de Vijver & Ligneul (2020), whereas other studies found intact RL even though paradigms were rather similar (Eppinger et al., 2008; Lighthall et al., 2018). In these paradigms, an image is presented and participants push one of two buttons to receive a positive, negative, or neutral reward. Then, the next image is presented randomly from a set of given images, and so on. The task in these studies is not very demanding, provoking only light deficits in both the young and old group. For this reason, Daniel et al. (2020) have proposed that deficits of older people are only found in demanding tasks.

Here, we used a more demanding paradigm, where the presentation of image *n* was not chosen randomly, as in the above studies, but depended on the choice made by the participants at image *n-1*. One image was a goal image, and participants were asked to find it as often as possible within a certain period of time (Fig. 5.2.1). Hence, the experiment mimics a navigation task, which is more realistic than the above paradigms and involves additional aspects such as a systematic exploration of the RL environment, linking states not only to actions but also to other states. To increase task demand, we tested not only a short (0.5s) but also a longer (6s) inter-stimulus intervals (ISI) between images. Various parameters were extracted from the participants' performance and a Q-learning model was fitted. To preface our results, with one exception, older participants' performance was comparable to that of younger controls.

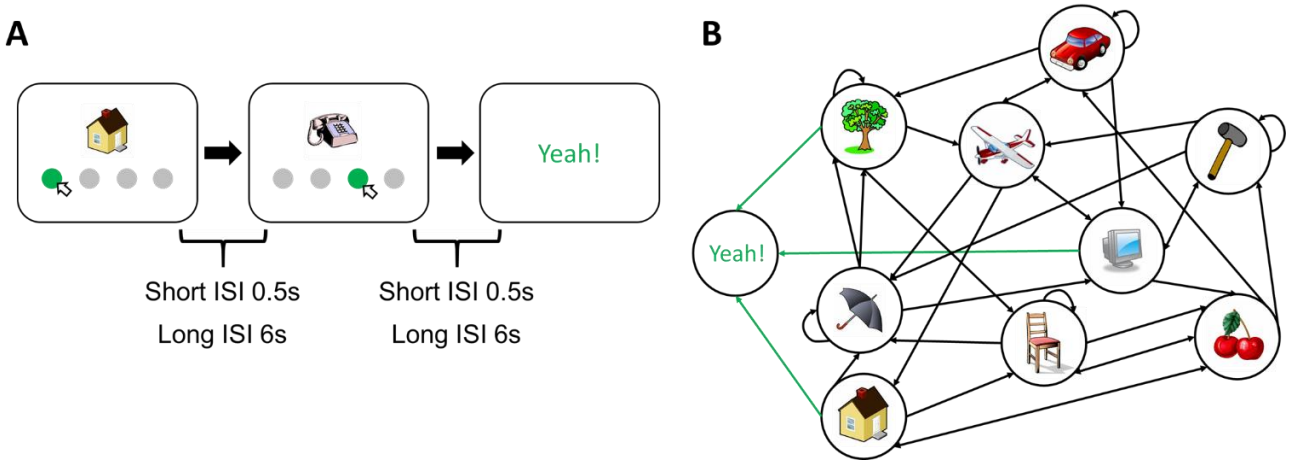

Figure 5.2.1. RL task. (A) An image was presented to the participants. Participants chose one of the disks below the image to proceed to the next image (state) until they found the goal state ("Yeah!"). Each participant performed under two different Inter-stimulus intervals (ISI) conditions, the short (0.5s) and long (6s). In experiment 1, participants had to reach the goal state as many times as possible within a limited duration (8 minutes for the short ISI and 30 minutes for the long ISI condition). In experiment 2, they had to find as many as possible goal states within 150 actions. Details can be found in the Methods and Materials. (B) Structure of the RL environment. Each image represents a state, and the direction of the arrow indicates the connection between different states. Green arrows indicate the path leading to the goal state. Importantly, the structure of the environment is very irregular in the sense that observers may go directly from image A to image B but not necessarily back. In all of the experiments, there are a total of nine states plus one goal state. As primary measures, we extracted 6 parameters including the number of episodes completed, the proportion of optimal actions, the improvement in both the number of episodes completed and the proportion of optimal actions, the learning rate and exploration rate (detailed in the Methods and Materials).

5.3 *Methods*

Experiment 1

Participants

Forty older healthy adults and thirty healthy young adults were recruited from the Free University of Tbilisi, Georgia. Detailed demographical information is presented in Table 5.3.1.

Task

Participants performed a reinforcement learning task (Fig. 5.2.1). Participants were presented with one image in the center of the screen, accompanied by four gray disks below. Clicking on one of the disks brought the participants to the next image. We call the clicks sometimes "actions" and the images "states" in accordance with RL terminology. Participants had no time limit for choosing a disk. The objective was to find a goal image, labeled with the word "Yeah!". Before the start of the experiment, all nine possible images were presented on the screen. The goal image was not presented but participants were informed about it. Observers initiated the experiment by clicking on a gray disk at the bottom of the screen.

We used two inter-stimulus intervals (ISI) of 0.5 seconds (short ISI condition) and 6 seconds (long ISI condition) between the participant's responses and the next state, respectively. The objective of the task was to reach the goal state as frequently as possible within a limited time of 8 minutes for the short ISI condition, and 30 minutes for the long ISI condition. The order of the two ISI conditions was counterbalanced among the participants, i.e., half began with the short ISI condition and the half started with the long ISI condition. There is one distinct transition matrix for each of the ISI condition, determining the transitions from one image to the next depending on the actions of the participants. Therefore, each participant is associated with two such matrices, which are identical across all participants.

Experiment 2

Participants

Twenty older healthy adults and twenty young healthy adults were recruited from the Free University of Tbilisi, Georgia. Individual who had previously taken part in the first experiment were no eligible. Detailed demographical information is presented in Table 5.3.1.

Experiment	Experiment 1		Experiment 2	
Group	Young	Old	Young	Old
Age	25.03±4.2	68.75±8.3	21.8±3.1	66.3±5.5
Gender	M=14, F=16	M=17, F=23	M=9, F=11	M=5, F=15
Education years	17.2±2.7	14.26±3.3	15.15±1.9	13.95±2.8
MoCa	NA	27.18±1.1	NA	27.75±1.5

Table 5.3.1. The demographic information of participants.

Task

Experiment 2 follows a similar procedure as experiment 1, with the difference that, instead of a fixed duration of 8 min, the number of trials was fixed at 150 for each of the ISI condition. Accordingly, the objective of the task was to reach the goal state as many times as possible within the given number of trials. Furthermore, the order of the two ISI conditions was fixed, with the long ISI condition always coming after the short ISI condition.

Immediately after the RL task, a memory task was given to the participants. In total, there were eighteen images. Twelve of these images were the same as in the RL task, the other six images were novel. For each of the images, participants were asked 1) to indicate whether the given image appeared in the RL task, and 2) to rate their confidence in question one, on a 4-Likert scale.

Behavioral analysis

Data pre-processing. In experiment 1, participants were instructed to reach the goal state as often as possible within 8min. Due to the nature of the task, different observers visited a different number of images ranging from 30 to 200 due to differences in decision-making and reaction times. To

ensure the comparability of the data among the participants, only the first 58 trials (the fifth percentile of the total number of trials among all participants) were used. Any trials exceeding this threshold were discarded. Four participants from the older population were removed from the study, as their total number of trials was lower than 59 trials. It is important to note that this pre-processing procedure only applied to experiment 1 as in experiment 2 the number of trials was fixed for all participants.

Behavioral performance. We determined the number of episodes completed and the proportion of optimal actions taken. The number of episodes completed refers to the number of times the participants reached the goal state. The proportion of optimal actions is the number of times a participant chose the action that optimally reduced the distance to the goal state from the current state divided by the total number of actions performed in the task. Furthermore, we assessed their improvement in performance by calculating the difference in the last and first 29 actions for the number of episodes completed and the proportion of optimal actions. A larger difference indicates better improvement, thereby implying enhanced learning. These four measures are our primary measures. Please note, these measures are not independent of each other. To gain further insights we used a number of secondary measures that are also not independent from the primary measures and can be seen as sub-measures.

First, the nine states of the environment were further categorized into adjacent and distant states. Adjacent states are states that were only one step away from the goal state. The six distant states were at least two steps away from the goal state. We analyzed performance separately for the two measures.

Second, to gain more insight into how performance progressed over time, we calculated the cumulative probability of optimal action across all trials. For each participant, we fitted a log function to estimate the intercept and the slope of the performance curve:

$$y = a + b * log\ (x)$$

x represents trial numbers, a and b are the intercept and the slope, respectively, y is the cumulative proportion of optimal actions in each trial.

Third, we tested perseveration behavior. It is essential to efficiently select the optimal actions for a given state in order to achieve superior performance. Choosing repeatedly a non-optimal action in a given state is suboptimal and may be related to perseverative behavior or memory deficits. We determined perseverations in two ways. First, we determined the average length of repeating action sequences. We averaged the length of these repetitions of actions across all episodes for each participant. For instance, an episode with the actions "1,2,3,1,2,3,4,2,1" has an average perseveration of two because the action sequence "1,2,3" appeared twice. Second, we calculated the proportion of repeated state-action pairs. In an episode, we extracted all the states and the corresponding actions to determine the probability that the same state-action pair has been visited within the same episode. In order to be considered optimal, a state should not be visited multiple times within an episode.

Fourth we determined the action entropy, measuring the randomness of the actions chosen. The theoretical max entropy is

$$E_{max} = -\sum(p_{max,j} * log_2 p_{max,j})$$

The probability distribution of completing each of the four actions was then determined for each state. The entropy of each state is

$$E_i = -\sum(p_{(i,j)} * log_2 p_{(i,j)}) \,/\, E_{max}$$

Here, *i* represents states, ranging from one to nine, while *j* represents actions, ranging from one to four. Averaging the action entropy across all states was computed for each participant. High action entropies indicate poor action choices (Sojitra et al., 2018).

Furthermore, we calculated the occupancy map for each state-action pair. Each cell in the matrices represents the number of times a participant engaged with a given state-action pair. Thus, a higher number in a cell indicates more frequent engagement with that specific state-action pair by the participant. Consequently, each participant will have two maps: one for the short ISI condition and one for the long ISI condition. To assess the similarity between these two maps for each participant, we used the cosine distance, calculated as follows:

$$\text{Cosine distance} = 1 - \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} * \sqrt{\sum_{i=1}^{n} B_i^2}}$$

where A_i and B_i are the elements of vectors A and B, respectively, with each element corresponding to the occupancy counts for a specific state-action pair. Given that there are nine different states and four possible actions from each state, the total number of unique state-action pairs is 36.

Computational modelling. A Q-learning model (Sutton & Barto, 2018) was used to quantify learning:

$$(S, A) \leftarrow Q(S, A) + \alpha(\Delta)$$

S represents the state of the current trial, *A* is the action taken in the current trial, Δ is the prediction error, Q is the Q-value for the given state and action, and α is the free parameter of the learning rate. Whenever a participant performed a new action, the Q value was updated by the prediction error, which is defined as the difference between the current reward and the expected reward:

$$\Delta = r + \max_i Q(S_{new}, A_i) - Q(S, A)$$

r stands for the reward, S_{New} is the new state after performing action *A* in the given state *S.* The probability of choosing an action is then determined by a soft-max rule:

$$q(S, A_i) = \frac{e^{\frac{Q(S,A_i)}{\tau}}}{\sum_j e^{\frac{Q(S,A_j)}{\tau}}}$$

q is the probability of choosing action *A* given state *S.* τ is the free parameter normally referred to as inverse temperature which corresponds to the exploration rate.

Statistical analysis. We conducted a two-by-two repeated measures ANOVA with age group (old and young) and ISI (short and long) as independent variables. We quantified the relationship between measurements using Pearson correlations, except for the relationship between measurements and Q-learning model parameters, which were calculated with Spearman's correlation due to the non-normal distribution of the parameters.

The primary parameters, focusing on performance, learning, and Q-learning model in RL, are presented in the main text. The secondary parameters, are found in the supplementary figures. All details are summarized in Table 5.4.S1, Table 5.4.S2, and Table 5.4.S3.

5.4 *Results*

Experiment 1

Effects of ISI and age. First, we examined whether performance was affected by age. Surprisingly, the number of episodes completed (Fig. 5.4.2A) and the proportion of optimal actions (Fig. 5.4.2B) did not differ significantly between young and old adults. Furthermore, the Bayes factor provide evidence in favor of the null hypothesis.

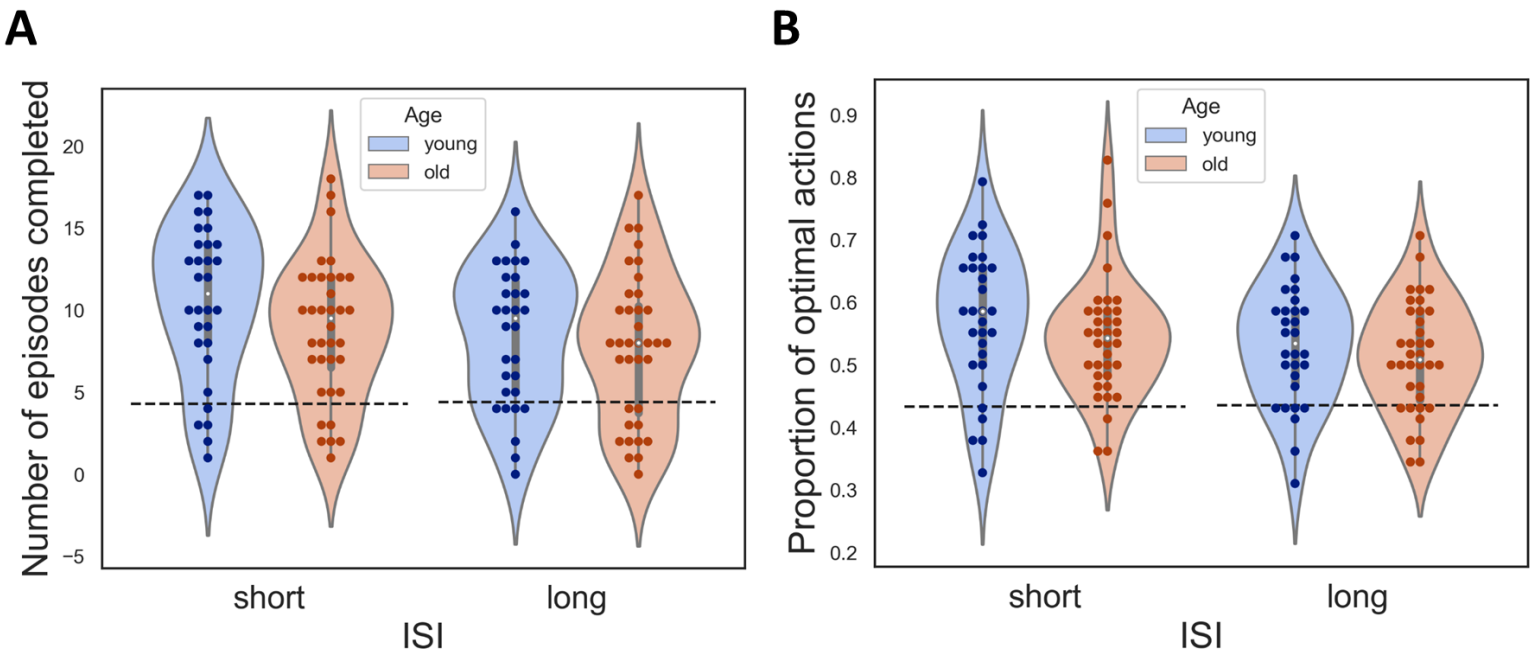

Figure 5.4.2. Performance in the RL task. (A) The number of episodes completed in each ISI condition for each age group. Dots indicate the performance of individual participants. (B) The proportion of optimal actions in each ISI condition for each age group. The dashed lines in the figures indicate baseline performance.

Improvement of performance. We next investigated the improvement of performance during the RL task by dividing the trials into two sets, the first 29 trials and the last 29 trials of each observer. There was a significant interaction between the two groups and ISI when it came to the number of episodes completed. A post-hoc test showed that the long ISI condition drove the interaction (Fig. 5.4.3A). A slight trend but no statistically significant improvement in performance was observed

(Fig. 5.4.3B). However, none of the Bayes factors for these results suggest support for the alternative hypothesis. For a closer examination, we fitted the accuracy data across trials with a log function which yielded two parameters, the slope and intercept. Both the intercept and the slope did not differ between the groups (Fig. 5.4.S2A).

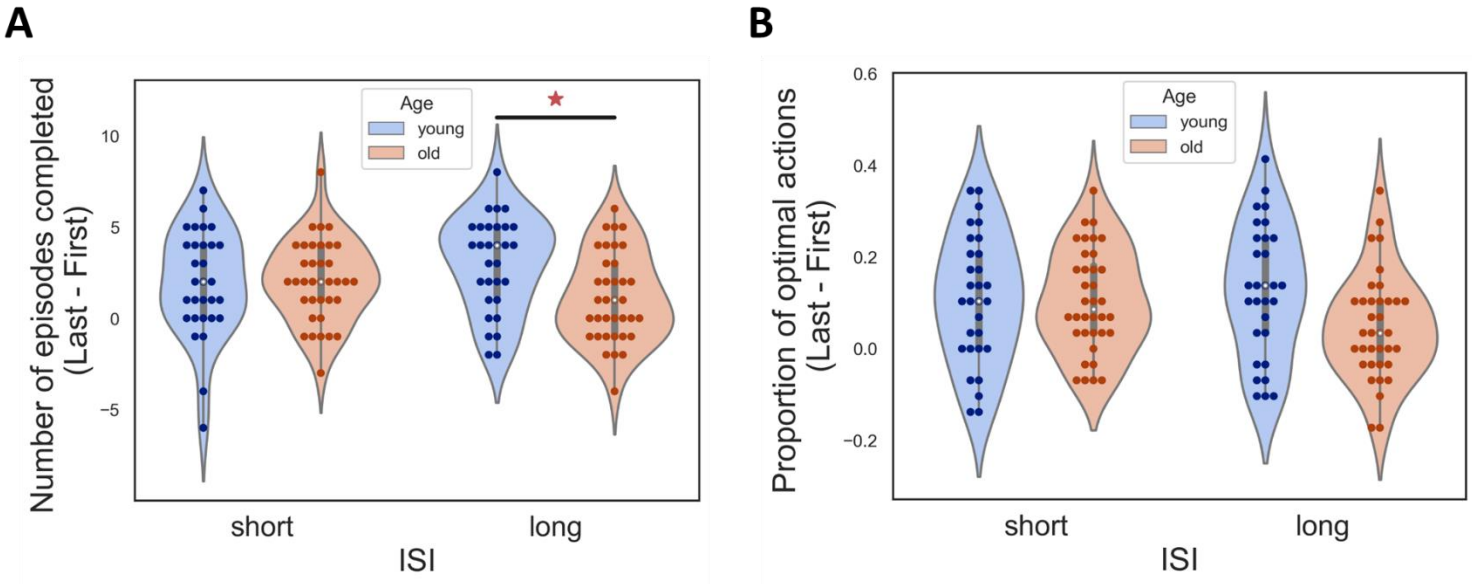

Figure 5.4.3. Improvement of performance. The improvement in the accuracy for each age group and ISI condition. (A) The improvement in number of episodes. *: $p < 0.05$ by post-hoc Tukey's test. (B) Improvement in proportion of optimal actions.

Perseveration behavior and action entropy. We assessed the extent of suboptimal action selection by quantifying the action entropy, which is a measure of the randomness of actions. Older adults had a lower improvement in action entropy compared to young adults (Fig. 5.4.S2B). We suggest below that the difference may be caused by memory deficits.

Additionally, we also examined the perseveration behavior, which is characterized by a persistent repetition of suboptimal actions. There were no significant differences between young and older adults (Fig. 5.4.S2C, left; Fig. 5.4.S2C, middle). Furthermore, the improvement in perseveration behavior was not significant across the age groups (Fig. 5.4.S2C, right).

Our findings indicate a difference in the improvement of action entropy between age groups, with the older group showing less improvement compared to the younger group, regardless of the ISI condition.

Q-learning. Both learning rates (Fig. 5.4.4A, top) and the inverse temperature (Fig. 5.4.4A, bottom) did not show significant differences between the two age groups. The learning rate reflects the speed of learning, and the inverse temperature reflects the randomness of choices or exploration rate. The results suggest that both groups revealed similar learning efficiency and exploration through the RL environment.

We next linked performance to the Q-learning model to study whether there are correlations between the model parameters and the improvement in performance. Interestingly, we found that only the learning rate in the long ISI condition positively correlated with improvement in performance (Fig. 5.4.4B, Bottom left). The inverse temperature was negatively correlated with improvement in performance merely in the short ISI condition (Fig. 5.4.4B, Top right), though the same negative trend was also present in the long ISI condition (Fig. 5.4.4B, Bottom right).

Effect size and statistical values can be found in Table 5.4.S1.

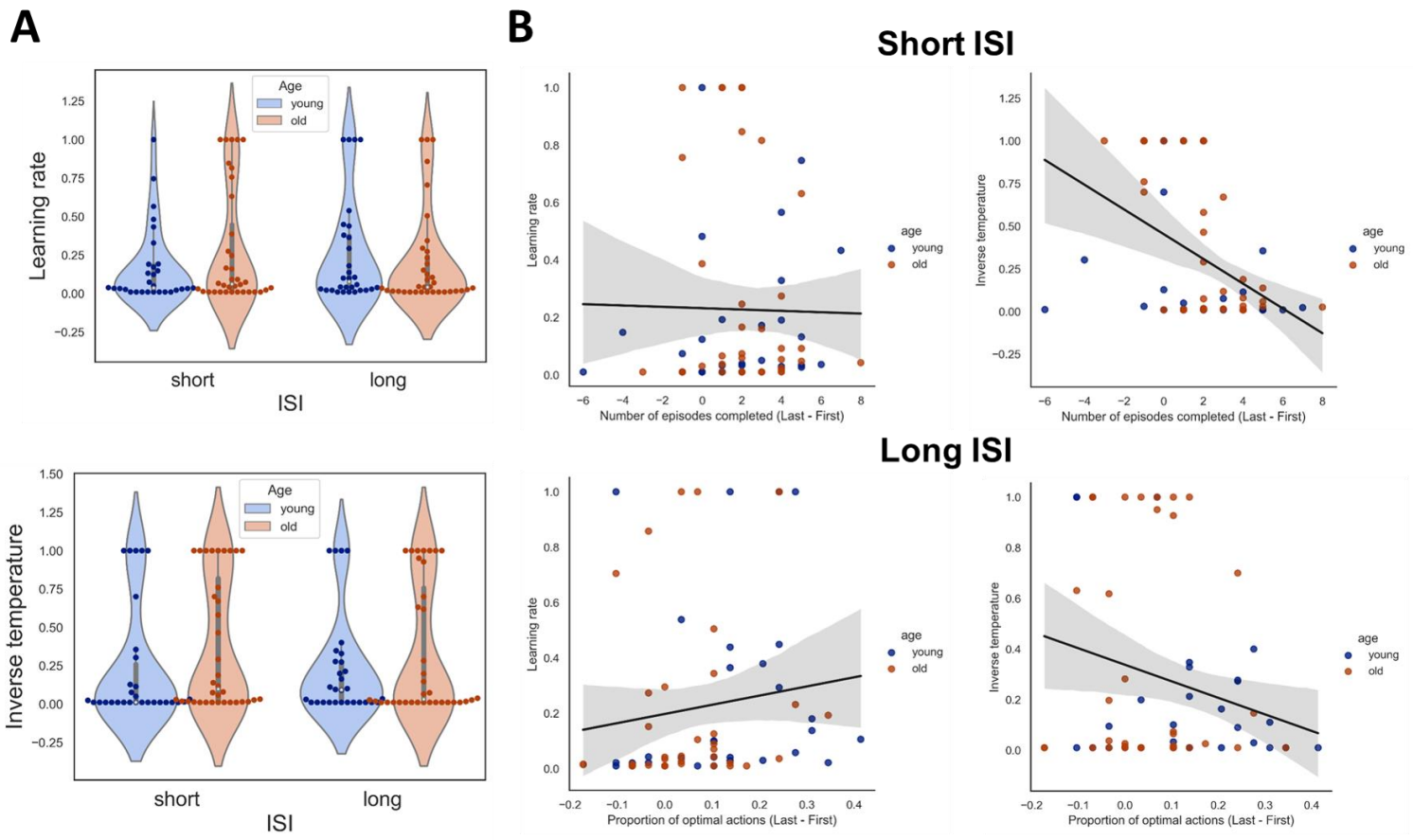

Figure 5.4.4. Q-learning model. The two parameters retrieved from the Q-learning model for each condition and for each age group: (A) Top: The learning rate; Bottom: The inverse temperature. The conventions are the same as in Fig. 2. (B) Correlation between the parameters and the improvement in performance. Top panel: The correlation between the improvement in performance and the learning rate (left) and the inverse temperature (right) in the short ISI condition. Bottom panel: The correlation between the improvement in performance and the learning rate (left) and the inverse temperature (right) in the long ISI condition. The black line is the linear regression fit of the parameters and performance improvement combined for both older and young adults.

Experiment 2

Effects of ISI and age. Contrary to experiment 1, a significant interaction of ISI and age was observed in the proportion of optimal actions (Fig. 5.4.5A, right), but a marginal difference in the number of episodes completed (Fig. 5.4.5A, left).

Moreover, there was a significant difference in improvement in performance in the long ISI between older and young adults (Fig. 5.4.5B). The Bayes factors also demonstrated moderate effects that support the alternative hypothesis.

The results indicate that only in the long ISI condition older and younger adults performed differently. We did not find significant differences in the behavioral patterns between the two groups, suggesting that similar strategies are used to perform the RL task (Fig. S1B, left, $t(38) = 1.19$, $p = 0.241$, Cohen's $d = 0.38$). Furthermore, we observed that the levels of similarity correlate with performance in both older and young participants. Hence, the measurements capture individual differences in task accuracy (Fig. 5.4.S1B, right $r_{young}(18) = 0.772$, $r_{old}(18) = 0.731$).

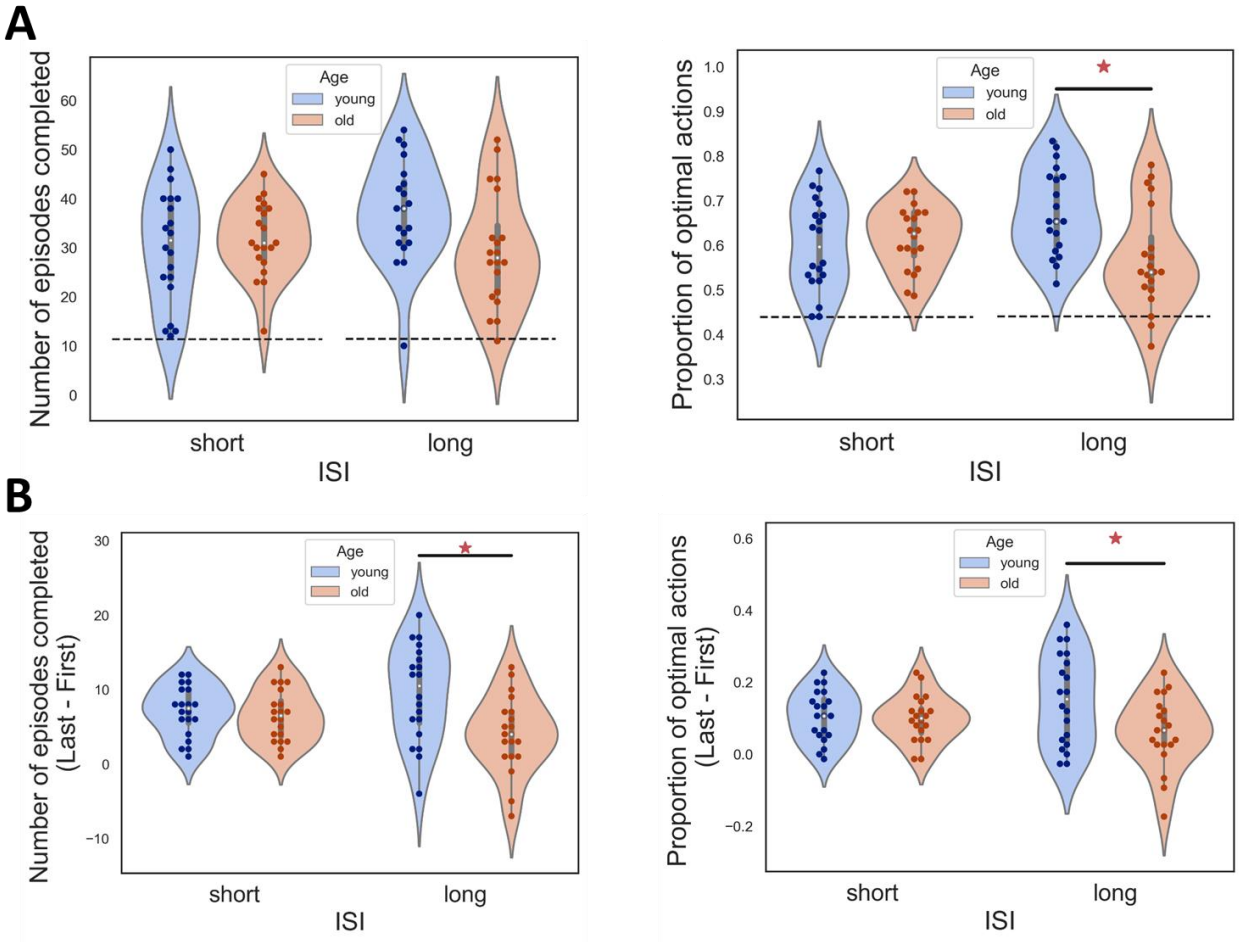

Figure 5.4.5. Performance in experiment 2. The conventions are the same as Fig. 5.4.1. (A) Left: The number of episodes completed in each ISI condition for each age group. Right: The proportion of optimal action in each ISI condition for each age group. *: $p < 0.05$ post-hoc Tukey's test. The dashed lines in the

figures indicate baseline performance (B) The improvement in the accuracy for each age group and ISI condition. Left: The improvement in number of episodes. Right: Improvement in proportion of optimal actions. *: $p < 0.05$ by post-hoc Tukey's test.

Questionnaire. The results of the questionnaire revealed that both older and young adults were able to accurately recognize the images used in the RL task, with no participants making mistakes. Next, we measured the confidence rating of their answers. The confidence ratings were based on a scale where a rating of 1 represented high confidence and a rating of 4 indicated high uncertainty. Interestingly, there was no significant difference in the confidence ratings between older and young adults in the adjacent states (Fig. 5.4.S3B, left). However, a clear main effect of age in the distant states (Fig. 5.4.S3B, right).

Q-learning. Similar to the results of experiment 1, both learning rates (Fig. 5.4.6A, top) and the inverse temperature (Fig. 5.4.6A, bottom) did not differ between the two age groups. This suggests that, while there may be some subtle differences in the learning processes of older and younger adults, overall there is not significant difference in the parameters used for Q-learning. However, despite the absence of group differences, the inverse temperature was found to be significantly correlated with accuracy (Fig. 5.4.6B, Top right) rather than learning rate (Fig. 5.4.6B, Top left). This indicates that while the learning rate may not have a direct effect on performance, the inverse temperature, which controls the exploration-exploitation trade-off, does affect the accuracy.

Effect size and statistical values can be found in Table 5.4.S2.

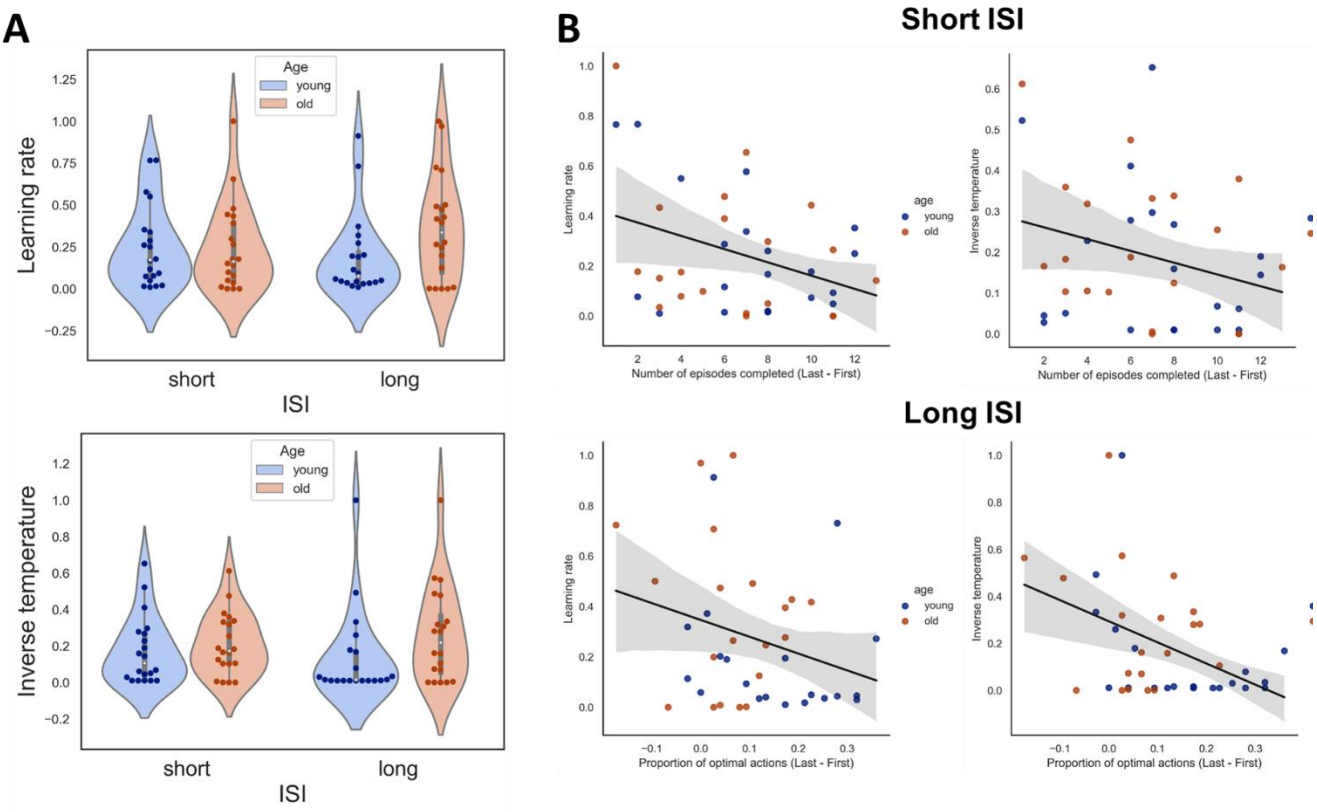

Figure 5.4.6. Q-learning model and the performance – experiment 2. The conventions are the same as in Fig. 5.4.4. A Q-learning model was fitted to the data and two parameters retrieved from the model for each condition and for each age group: (A) Top: The learning rate; Bottom: The inverse temperature. (B) The correlation between the parameters and the proportion of optimal actions. Top panel: The correlation between the proportion of optimal actions and the learning rate (left) and the inverse temperature (right) in the short ISI condition. Bottom panel: The correlation between the proportion of optimal actions and the learning rate (left) and the inverse temperature (right) in the long ISI condition.

5.5 *Discussion*

In the current study, we aimed to investigate whether older adults exhibit deficits in RL. In experiment 1, older adults performed as well as young adults with the number of episodes completed, proportion of optimal actions, improvement in proportion of optimal actions, learning rate and the exploration rate. In experiment 2, we found significant differences in the long ISI condition for improvement in accuracy and the secondary measures derived from this measure. Overall, the results of our study are consistent with previous studies (Lighthall et al., 2018; Pietschmann et al., 2011) but not others (Daniel et al., 2020). Hence, RL is largely intact in older observers with our paradigm.

Moderate to large effects were observed in the measurements reflecting the performance differences between the two age groups in the long ISI condition. There are several explanations: a genuine RL deficit, a working memory deficit, or stronger fatigue. First, none of the parameters of the Q-learning model was abnormal in the older participants, including learning and exploration rates, which speaks rather against an RL deficit.

Second, in the memory questionnaire, older adults demonstrated less confidence in recognizing distant states. In addition, the improvement of action entropy was less pronounced. These findings may speak indeed to a slight memory deficit. Lighthall et al. (2018), employed a RL task also with short and long ISIs. Although the authors did not observe behavioral differences between the two age groups for either ISI condition, they found a significant difference in the hippocampal activity pattern between the age groups in the long ISI condition. Potentially the higher memory load in the long ISI condition leads to differences in hippocampal activation, but does not manifest on the behavioral level because of the simpler task. Furthermore, we did not observe significant differences in the similarity of occupancy maps across the two ISI conditions between the age groups, indicating that both groups performed in a similar manner. More interestingly, this measure can capture individual differences in accuracy.

Third, it is well known that older people fatigue more quickly than younger ones (Enoka & Duchateau, 2016). Hence, instead of memory load, higher fatigue in the older population may explain the results in the long ISI condition. However, our analysis of adjacent and distant states shows that only the distant states exhibited differences in terms of the proportion of optimal actions,

indicating that the older adults were still able to locate the correct actions when the goal state was close enough. Hence, a memory deficit seems to be the best explanation at the moment.

As mentioned in the introduction, Daniel et al. (2020) claimed that older people perform worse than younger ones in demanding but not simpler tasks. The study had an easy and a harder condition and found no performance differences in the easy condition but a trend for a group comparison in the hard condition. However, the conclusion is potentially not valid because of a ceiling effect in the easy condition (performance between younger and older controls: 97% vs. 94%). Potentially, there is a group effect also in the easy condition but definitely in the harder condition. Since task demand seems not to be the crucial aspect, there must be other reasons for the differences in the paradigms.

One of the limitations in the ageing field is that the older population exhibits a high variance due to differences in ageing. In addition, the mean age is often different. Small sample sizes may lead naturally to sampling biases, which may lead to different outcomes. In addition, there are demoscopical, socio-economical, and cultural differences. Hence, it may simply be the case that our and others null results come from a too "healthy" and/or still too "young" population (our mean age was 68 and 66 years and in Daniel et al. (2020) it was 70 years). Thus, not the paradigms and differences between paradigms are key but sampling of the population.

Next to performance, learning and exploration rate, we also tested for perseverations, a behavioral pattern where individuals continue to perform repeated action sequences regardless of whether they are optimal or not. We did not find an increased perseveration rate in older adults, contrary to previous results, for example, in the Wisconsin Card Sorting test (Shaqiri et al., 2019). With the very same paradigm used in this study, we tested schizophrenia patients and found evidence for perseverations compared to age-matched controls. Hence, our paradigm is sensitive to perseverations. Perseveration of schizophrenia patients are often attributed to abnormal dopamine levels leading to suboptimal action selection (Durstewitz & Seamans, 2008). In RL models, dopamine levels are classically related to the learning rate (Sutton & Barto, 2018) and prediction error (Schultz et al., 1997; Wise & Rompre, 1989). If this were all true, our results indicate an intact dopaminergic system.

In summary, our findings indicate that reinforcement learning can remain relatively preserved during ageing. This holds true for a large range of outcome measures we determined including learning and exploration rate in classic Q-learning models. We emphasize the importance to publish such null results. Suppressing null results may, otherwise, create the impression that older people are deficient in almost all paradigms, which does not seem to be true.

Acknowledgements

This project was funded by the Swiss National Science Foundation (SNF) NCCR project Synapsy: The Synaptic Bases of Mental Diseases and the Sinergia project “Learning from Delayed and Sparse Feedback”. We would like to thank Aaron Clarke for his significant contributions to both the model and the experimental design.

Disclosure Statement

There are no conflicts of interest to disclose. The manuscript is original, has not been previously published and is not currently under consideration for publication elsewhere.

Supplementary figures

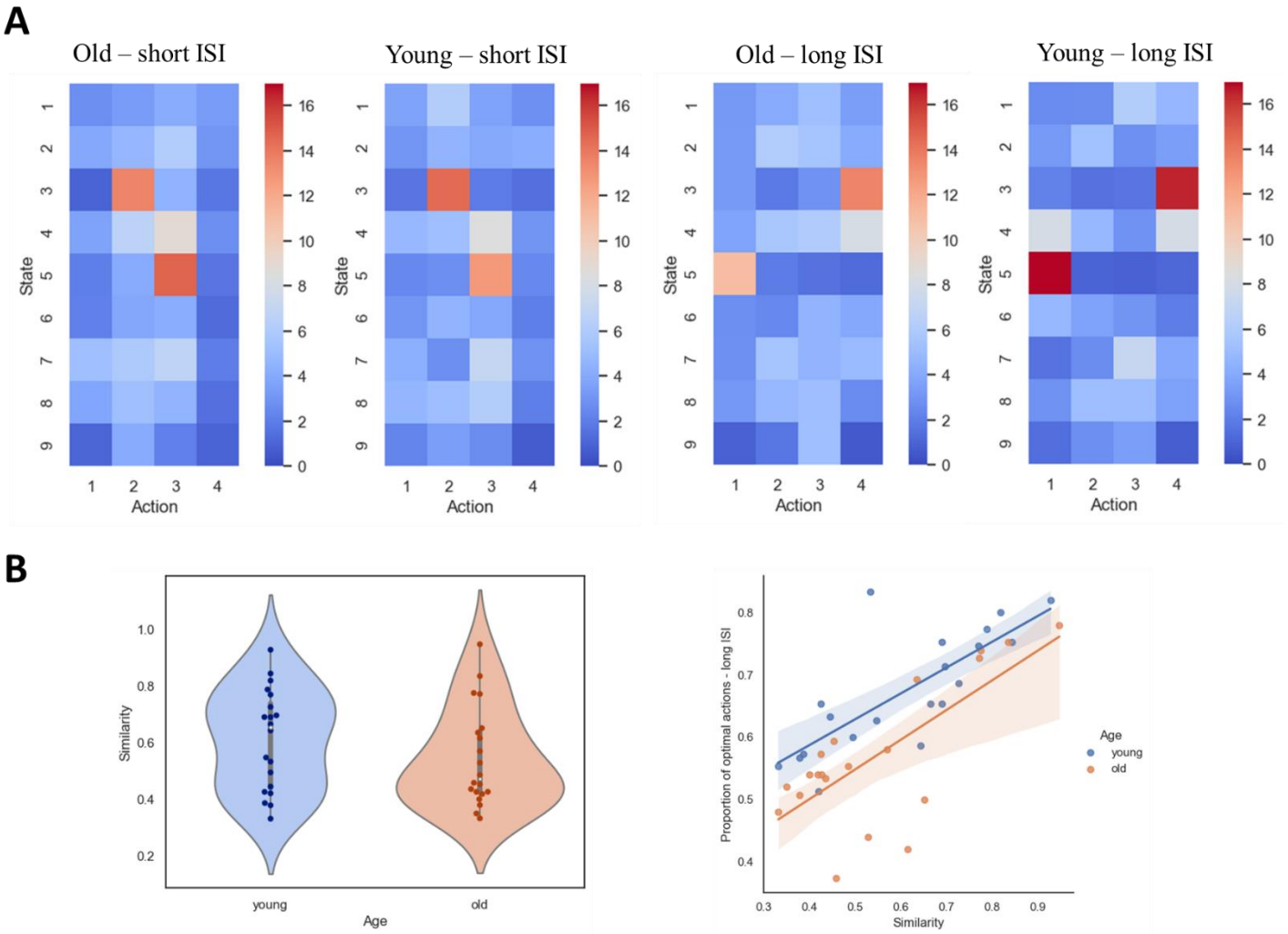

Figure 5.4.S1. The similarity between the two ISI conditions predicts RL performance. (A) The average occupancy map for both age groups. The values in each cell represent the average counts for performing specific state-action pairs. (B) Left: The similarity of the maps between the two ISI conditions for each age group. Right: The correlation between the proportion of optimal actions and the levels of similarity.

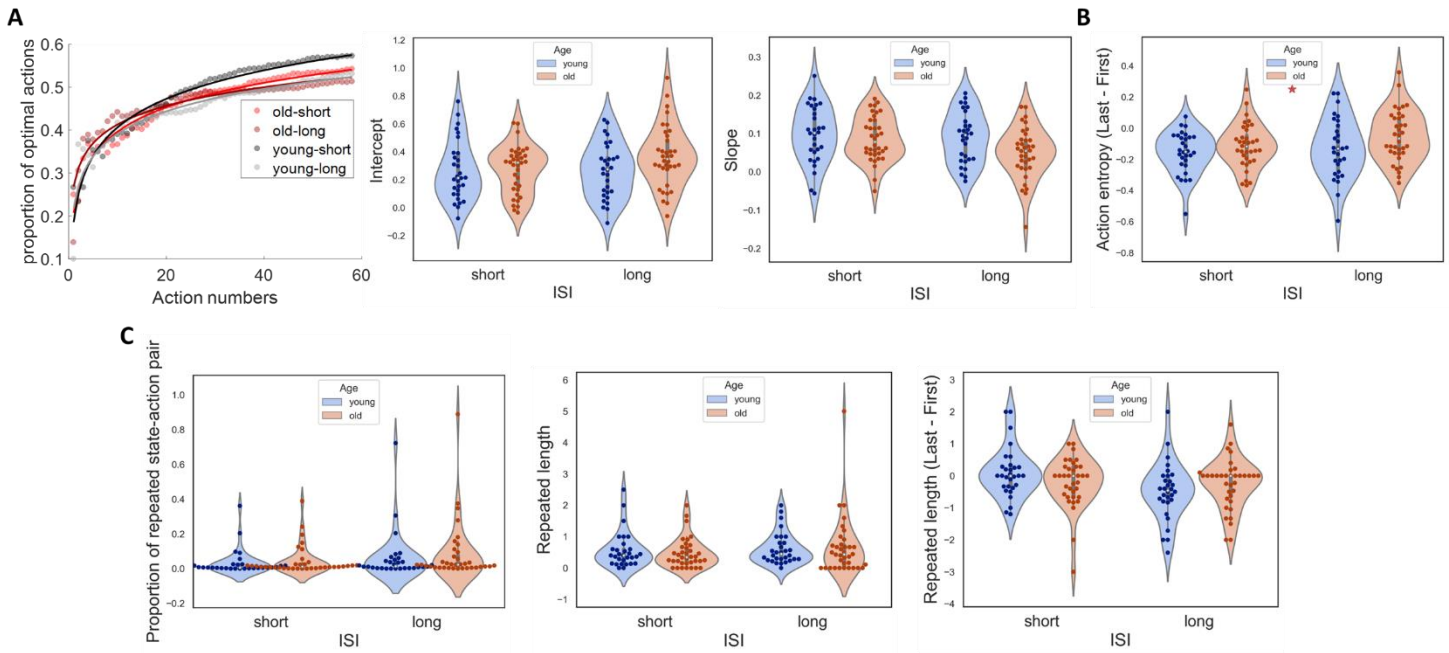

Figure 5.4.S2. Perseveration behavior and action entropy for experiment 1. (A) Left: The cumulative proportion of optimal actions across time. The two ISI conditions and age groups were plotted separately. Each dot represents the average cumulative proportion of optimal actions across the participants in each age group. The solid line shows the fit of the log function. Right: The two bar graphs depict the slopes and intercepts of the fitted log function. (B) The improvement in action entropy. A main effect of age is presented. *: $p < 0.05$. (C) Left: The proportion of repeated state-action pair of each age group in each ISI condition. Middle: The repeated action length of each age group in each ISI condition. Right: The improvement in the repeated action length.

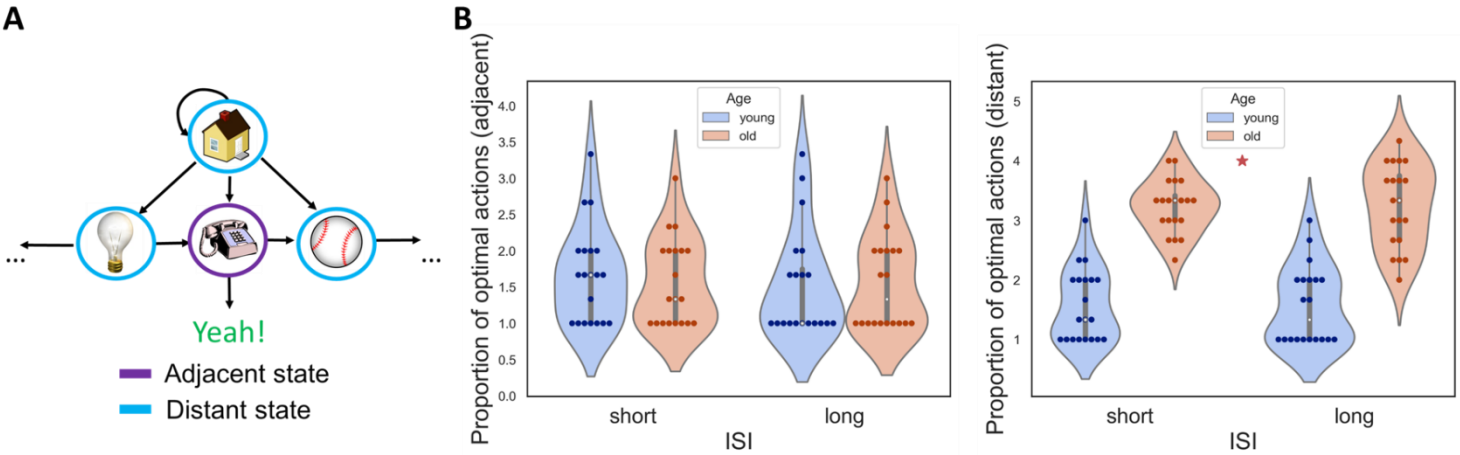

Figure 5.4.S3. Secondary measurements for experiment 2. (A) The adjacent state is indicated in purple, and refers to a state that is directly connected to the goal state. Distant states, indicated in blue, are those that require multiple steps to reach to the goal state. (B) Confidence ratings of the questionnaire. Left: The average confidence ratings in adjacent states for each age group in each ISI condition. Right: The average confidence ratings in distant states for each age group in each ISI condition.

Primary parameters – experiment 1	**Short ISI**	**Long ISI**
Number of episodes completed (2A)	age: $F(1,64) = 1.98$, $p = 0.16$, partial $\eta^2 = 0.03$ interaction: $F(1,64) = 0.44$, $p = 0.51$, partial $\eta^2 = 0.007$ $BF_{10\_age} = 0.48$	
Proportion of optimal actions (2B)	age: $F(1,64) = 1.83$, $p = 0.18$, partial $\eta^2 = 0.03$ interaction: $F(1,64) = 0.14$, $p = 0.71$, partial $\eta^2 = 0.002$ $BF_{10\_age} = 0.43$	
Improvement in number of episodes completed (3A)	Age: $F(1,64) = 3.66$, $p = 0.06$, partial $\eta^2 = 0.054$ interaction: $F(1,64) = 5.4$, **$p = 0.02$**, partial $\eta^2 = 0.078$ $BF_{10\_age} = 0.84$	
	ns	post-hoc test **$p_{tukey} = 0.018$** Cohen's d = 0.74
Improvement in proportion of optimal actions (3B)	age: $F(1,64) = 3$, $p = 0.09$, partial $\eta 2 = 0.045$ interaction: $F(1,64) = 3.1$, $p = 0.08$, partial $\eta 2 = 0.046$ $BF_{10\_age} = 0.68$	
Q-learning-alpha (4A)	age: $F(1,64) = 0.39$, $p = 0.54$, partial $\eta^2 = 0.006$ interaction: $F(1,64) = 1.77$, $p = 0.19$, partial $\eta^2 = 0.027$ $BF_{10\_age} = 0.24$	
Q-learning-tau (4A)	age: $F(1,64) = 2.98$, $p = 0.09$, partial $\eta^2 = 0.044$ interaction: $F(1,64) = 0.07$, $p = 0.79$, partial $\eta^2 = 0.001$ $BF_{10\_age} = 0.68$	
Secondary parameters – experiment 1		
Intercept of log fitting (S2A)	$F(1,64) = 0.65$, $p = 0.42$, partial $\eta^2 = 0.01$, $BF_{10\_age} = 0.58$	
Slope of log fitting (S2A)	$F(1,64) = 1.15$, $p = 0.29$, partial $\eta^2 = 0.02$, $BF_{10} = 1.09$	
Improvement in action entropy (S2B)	age: $F(1,64) = 5.21$, **$p = 0.026$**, partial $\eta^2 = 0.075$; interaction: $F(1,64) = 0.37$, $p = 0.55$, partial $\eta^2 = 0.006$ $BF_{10\_age} = 1.09$	
Proportion of repeated-action pair (S2C)	age: $F(1,64) = 0.06$, $p = 0.81$, partial $\eta^2 = 0.0009$ interaction: $F(1,64) = 0.39$, $p = 0.53$, partial $\eta^2 = 0.006$ $BF_{10\_age} = 0.23$	
Repeated action length (S2C)	age: $F(1,64) = 0.59$, $p = 0.45$, partial $\eta^2 = 0.009$ interaction: $F(1,64) = 0.09$, $p = 0.77$, partial $\eta^2 = 0.001$ $BF_{10\_age} = 0.27$	

Table 5.4.S1. Summary of parameters for experiment 1.

Primary parameters – experiment 2	**Short ISI**	**Long ISI**
Number of episodes completed (5A)	age: $F(1,38) = 1.55$, $p = 0.22$, partial $\eta^2 = 0.04$ interaction: $F(1,38) = 4.84$, **$p = 0.03$**, partial $\eta^2 = 0.11$ $BF_{10} = 0.55$	
	ns	ns
Proportion of optimal actions (5A)	age: $F(1,38) = 3.55$, $p = 0.07$, partial $\eta^2 = 0.08$ interaction: $F(1,38) = 9.77$, **$p = 0.003$**, partial $\eta^2 = 0.2$ **$BF_{10} = 5.9$**	
	ns	post-hoc test **$p_{tukey} = 0.005$** Cohen's d = -1.1
Improvement in number of episodes completed (5B)	age: $F(1,38) = 10.1$, **$p = 0.003$**, partial $\eta^2 = 0.21$ interaction: $F(1,38) = 4.37$, **$p = 0.043$**, partial $\eta^2 = 0.1$ **$BF_{10\_age} = 3.29$**	
	ns	post-hoc test **$p_{tukey} = 0.003$** Cohen's d = -1.14
Improvement in proportion of optimal actions (5B)	age: $F(1,38) = 5.1$, **$p = 0.03$**, partial $\eta^2 = 0.12$ interaction: $F(1,38) = 4$, $p = 0.052$, partial $\eta^2 = 0.095$ $BF_{10\_age} = 1.41$	
	ns	post-hoc test **$p_{tukey} = 0.018$** Cohen's d = -0.95
Q-learning-alpha (6A)	age: $F(1,38) = 0.04$, $p = 0.84$, partial $\eta^2 = 0.001$ interaction: $F(1,38) = 0.26$, $p = 0.62$, partial $\eta^2 = 0.007$ $BF_{10} = 0.13$	
Q-learning-tau (6A)	age: $F(1,38) = 1$, $p = 0.32$, partial $\eta^2 = 0.026$ interaction: $F(1,38) = 0.034$, $p = 0.86$, partial $\eta^2 = 0.0009$ $BF_{10} = 0.07$	
Secondary parameters – experiment 2		
Memory questionnaire – adjacent (S3)	age: $F(1,38) = 0.03$, $p = 0.86$, partial $\eta^2 = 0.0008$ interaction: $F(1,38) = 0.6$, $p = 0.44$, partial $\eta^2 = 0.016$ $BF_{10} = 0.05$	
Memory questionnaire – distant (S3)	age: $F(1,38) = 110.37$, **$p < 0.001$**, partial $\eta^2 = 0.74$ interaction: $F(1,38) = 0.006$, $p = 0.94$, partial $\eta^2 = 0.0004$ **$BF_{10\_age} = 9.04x10^9$**	

Table 5.4.S2. Summary of parameters for experiment 2.

Parameter 1	**Parameter 2**	**Short ISI**	**Long ISI**
Improvement in number of episodes completed	Q-learning alpha (4B)	$r(64) = 0.4$, $p < 0.001$	ns
Improvement in number of episodes completed	Q-learning tau (4B)	$r(64) = -0.4$, $p < 0.001$	ns
Proportion of optimal actions	Q-learning alpha (6B)	ns	ns
Proportion of optimal actions	Q-learning tau (6B)	$r(38) = -0.38$, $p = 0.015$	$r(38) = -0.5$, $p = 0.008$

Table 5.4.S3. Summary of all correlation measurements.

5.6 *References*

Anguera, J. A., Reuter-Lorenz, P. A., Willingham, D. T., & Seidler, R. D. (2011). Failure to Engage Spatial Working Memory Contributes to Age-related Declines in Visuomotor Learning. *Journal of Cognitive Neuroscience*, *23*(1), 11–25. https://doi.org/10.1162/jocn.2010.21451

Daniel, R., Radulescu, A., & Niv, Y. (2020). Intact Reinforcement Learning But Impaired Attentional Control During Multidimensional Probabilistic Learning in Older Adults. *The Journal of Neuroscience*, *40*(5), 1084–1096. https://doi.org/10.1523/JNEUROSCI.0254-19.2019

Durstewitz, D., & Seamans, J. K. (2008). The Dual-State Theory of Prefrontal Cortex Dopamine Function with Relevance to Catechol-O-Methyltransferase Genotypes and Schizophrenia. *Biological Psychiatry*, *64*(9), 739–749. https://doi.org/10.1016/j.biopsych.2008.05.015

Enoka, R. M., & Duchateau, J. (2016). Translating Fatigue to Human Performance. *Medicine & Science in Sports & Exercise*, *48*(11), 2228–2238. https://doi.org/10.1249/MSS.0000000000000929

Eppinger, B., Kray, J., Mock, B., & Mecklinger, A. (2008). Better or worse than expected? Aging, learning, and the ERN. *Neuropsychologia*, *46*(2), 521–539. https://doi.org/10.1016/j.neuropsychologia.2007.09.001

Lighthall, N. R., Pearson, J. M., Huettel, S. A., & Cabeza, R. (2018). Feedback-Based Learning in Aging: Contributions and Trajectories of Change in Striatal and Hippocampal Systems. *The Journal of Neuroscience*, *38*(39), 8453–8462. https://doi.org/10.1523/JNEUROSCI.0769-18.2018

Pietschmann, M., Endrass, T., Czerwon, B., & Kathmann, N. (2011). Aging, probabilistic learning and performance monitoring. *Biological Psychology*, *86*(1), 74–82. https://doi.org/10.1016/j.biopsycho.2010.10.009

Salthouse, T. A. (2009). Decomposing age correlations on neuropsychological and cognitive variables. *Journal of the International Neuropsychological Society*, *15*(5), 650–661. https://doi.org/10.1017/S1355617709990385

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Shaqiri, A., Pilz, K. S., Cretenoud, A. F., Neumann, K., Clarke, A., Kunchulia, M., & Herzog, M. H. (2019). No evidence for a common factor underlying visual abilities in healthy older people. *Developmental Psychology*, *55*(8), 1775–1787. https://doi.org/10.1037/dev0000740

Sojitra, R. B., Lerner, I., Petok, J. R., & Gluck, M. A. (2018). Age affects reinforcement learning through dopamine-based learning imbalance and high decision noise—Not through Parkinsonian mechanisms. *Neurobiology of Aging*, *68*, 102–113. https://doi.org/10.1016/j.neurobiolaging.2018.04.006

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (Second edition). The MIT Press.

van de Vijver, I., & Ligneul, R. (2020). Relevance of working memory for reinforcement learning in older adults varies with timescale of learning. *Aging, Neuropsychology, and Cognition*, *27*(5), 654–676. https://doi.org/10.1080/13825585.2019.1664389

Wise, R. A., & Rompre, P.-P. (1989). Brain Dopamine and Reward. *Annual Review of Psychology*, *40*(1), 191–225. https://doi.org/10.1146/annurev.ps.40.020189.001203

6. Study IV: Behavioral and Neural Markers of Social Dominance: A Female-Focused Perspective

In the last study, we aimed to identify universal behavioral and neural markers for social dominance. Given that da Cruz et al. (2018) exclusively examined male participants, our focus was specifically on the female population. We thus adopted the same procedure as in da Cruz et al. (2018), which combined an emotion discrimination task with EEG recording. Additionally, we included a resting state EEG session to ascertain any differences between the high and low dominance groups without involvement of task. While there was a trend indicating dominant females responded faster, this difference in response times was not statistically significant. Surprisingly, we observed a similar EEG N2/P2 component in the female participants as in the male population, suggesting that this could be a general neural marker for social dominance. Furthermore, resting state EEG data showed higher connectivity in the high dominant group compared to the low dominance group, suggesting intrinsic differences in brain network connectivity between the two groups, independent of task demands.

Behavioral and Neural Markers of Social Dominance: A Female-Focused Perspective

Wei-Hsiang Lin[1], Janir Ramos da Cruz[1,3], Carmen Sandi[2] & Michael H. Herzog[1]

1. Laboratory of Psychophysics, Brain Mind Institute, School of Life Sciences, Swiss Federal Institute of Technology Lausanne (EPFL), CH-1015 Lausanne, Switzerland
2. Laboratory of Behavioral Genetics, Brain Mind Institute, School of Life Sciences, Swiss Federal Institute of Technology Lausanne (EPFL), CH-1015 Lausanne, Switzerland
3. Wyss Center for Bio and Neuroengineering, Geneva, Switzerland

6.1 *Abstract*

Social interactions are crucial in human life, with social dominance being a key factor. We have previously shown that dominant males were faster in decision-making tasks than non-dominant ones, even in the absence of a social context such as competition. In addition, dominant males had a much higher N2/P2 EEG component, which is associated with the allocation of resources for decision-making. The EEG component provide evidence of an inherent trait and, thus, may serve as a neuromarker for social dominance. In our current study, we show that dominant females also have a larger N2/P2 EEG component compared to less dominant females. The latency and amplitude of this component closely mirrored those observed in the previous study with males. Additionally, we analyzed resting-state EEG and observed a higher whole-brain network density in the dominant females compared to less dominant ones, potentially representing a marker for social dominance, detectable at baseline (i.e., without a task challenge).

6.2 *Introduction*

Social dominance is a fundamental component of social groups, often manifesting in deep hierarchical structures observed within diverse socio-political and economic systems (Pratto et al., 1994; Redhead et al., 2019).

Classically, social dominance has been investigated in paradigms where, for example, two or more participants compete for different ranks, and performance becomes the defining criterion for the dominance level. In these studies, elevated activity in the dorso-lateral prefontal cortex (dlPFC) and the striatum correlate with superior hierarchical positions among participants (Ligneul et al., 2016; Santamaría-García et al., 2014; Zhou et al., 2018; Zink et al., 2008). A key insight from these studies is that social dominance is a relative scale, fundamentally relying on the comparison with others.

However, several studies have suggested that social dominance is a distinct personality trait (Hall et al., 2005; Schutz, 1958). Dominant individuals often exhibit characteristic behaviors, including extended speaking durations (Mast, 2002) and a higher perception of competence (Anderson and Kilduff, 2009). These findings imply that intense social interaction or competition might not be prerequisites for assessing social dominance; instead, it may be an intrinsic trait. This hypothesis received support from our previous study (da Cruz et al., 2018) where male participants were faster in complex decision making, but not in simple reaction, tasks than less dominant males. In addition, a higher N2/P2 EEG component was observed in high dominant males when performing an emotion discrimination task. This indicates that dominant males allocate more resources when making decisions (Mudar et al., 2016).

Social dominance is predominantly studied in male participants (da Cruz et al., 2018; Ligneul et al., 2016). While some studies have suggested an interplay between levels of social dominance and gender in stress behavior (Karamihalev et al., 2020) and in political attitudes (Pratto et al., 1997), other research, such as Scheggia et al. (2022), reports no gender differences in social-decision making scenarios. Notably, these prior investigations neither systematically explored whether gender interacts with social dominance in certain behaviors, nor incorporated neuroimaging to identify potential neuromarkers for social dominance. To address this gap, we adopted the procedure from da Cruz et al. (2018) and tested exclusively female participants to

investigate: 1) if dominant females respond more quickly to decision-making tasks, and 2) whether the N2/P2 EEG component, previously identified in males, is also evident in females. Additionally, we included resting-state EEG recordings to explore potential differences in brain network connectivity in the absence of task engagement.

6.3 *Methods*

Participants

We recruited twenty-six female participants from both the Swiss Federal Institute of Technology in Lausanne (EPFL) and the University of Lausanne (UNIL). Participants completed a standardized handedness questionnaire (Oldfield 1971). We then evaluated their visual acuity in both eyes using the Freiburg Visual Acuity Test (Bach 1996), stipulating a minimum acuity of 1.0 for both eyes as a requisite for participation. As compensation for their participation, they received 30 CHF per hour for EEG sessions, and 25 CHF per hour for behavioral sessions. To ensure uniform cortisol levels, EEG experiment sessions were scheduled between 1 PM and 7 PM.

Personality measurements

Before the experiment, participants completed a series of four questionnaires to assess social dominance motivation, state-trait anxiety, and menstrual cycle status. The questionnaires included the Personality Research Form dominance (PRF-D; Jackson, 1974), two aspects of Speilberger's State-Trait Anxiety Inventory (Spielberger, 1983), and a menstrual cycle status questionnaire. Each was administrated individually through the online experimental platform, Gorilla (Anwyl-Irvine et al., 2020). To negate potential order effects, we randomized both the questionnaire sequence and the item order within each.

We gauged participants' social dominance motivation using the PRF-D, which they completed at least three days before the experiment. This 16-item form contains both positive and negative items, utilizing a true/false response format. Speilberger's State-Trait Anxiety Inventory (STAI) assesses both trait (STAI-T) and state (STAI-S) anxiety via two separate sections, each containing 20 items. Participants respond on a 4-point Likert scale, ranging from 1 (completely disagree) to 4 (completely agree). Scores can vary from 20 (minimal anxiety) to 80 (extreme anxiety).

Lastly, we presented a two-part questionnaire about menstrual cycle status. The first question discerned if participants were on medications or treatments influencing hormone levels.

The second estimated the time since their last menstrual cycle, offering responses ranging from less than a week to three weeks.

EEG Experiment

The present study was conducted based on the protocol of Experiment 5 of da Cruz et al. (2018). We employed 40 emotional faces, both male and female, exhibiting happiness, sadness, anger, and neutrality. Faces displaying happiness and sadness were sourced from Ekman and Friesen's Pictures of Facial Affect Series (Ekman and Friesen 1976). In contrast, the angry and the neutral expressions were derived from FACES (Ebner et al. 2010), Radboud Faces Database (Langner et al. 2010) and the Karolinska Directed Emotional Faces (Lundqvist et al. 1998). Face were presented on a grey background and the luminance was below 1 cd/m^2. Participants were positioned 50 cm away from the monitor inside a dimly lit Faraday cage. We utilized The Eye Tribe © eye tracker to monitor gaze consistently, and a chin rest was used to stabilize participants' heads.

At the start of each trial, participants were asked to fixate a cross at the center of the screen, with an interval varying randomly between 0.5 to 1.5 seconds. A face image was then presented for 0.1 seconds at 26° either to the left or right of the center. Participants had up to 3 seconds to identify the depicted emotion using one of two buttons, each held in a different hand. Failure to respond triggered a short buzzer. The trial was repeated at the end of the block.

The experiment comprised two separate conditions with their order randomized for participants. Condition 1 utilized happy and sad faces (Happy vs. Sad) as stimuli, whereas condition 2 featured angry and neutral expressions (Angry vs. Neutral). The association between response side and emotional valence was counterbalanced among participants. Each condition started with a 10-trial practice round, succeeded by five experimental blocks comprising 80 trials each. Instructions were displayed on the screen before each condition began. Participants were instructed to respond as quickly and accurately as possible, and maintaining their gaze on the central fixation cross.

At the end of the experiment, we captured a 5-minute resting state EEG with participants keeping their eyes closed.

EEG Recording and Data Pre-processing

EEG data was acquired using an Active Two system (BioSemi, The Netherlands) equipped with 128 Ag-AgCl sintered active electrodes, all referenced to the common mode sense (CMS) electrode. The cap's size and placement were individually adjusted to ensure a proper fit for each participant. The cap was positioned so that the A1 electrode came to lie midway between the inion and the nasion, ensuring a balanced electrode coverage of frontal and occipital regions. Electrooculogram (EOG) recording were captured using electrodes placed 1 cm above and below the right eye and 1 cm lateral to the outer canthus. The EEG recording sampling rate was 2048 Hz. Data was downsampled offline to 512 Hz and underwent a comprehensive pre-processing pipeline based on the protocol by da Cruz et al. (2018). This included bandpass filtering from 1 to 50 Hz via a 3rd order Butterworth filter, line-noise removal with CleanLine (www. nitrc.org/projects/cleanline), re-referencing to the bi-weight mean estimate of all channels (Hoaglin et al. 1982), elimination of bad channels followed by 3D spline interpolation, removal of bad epochs, and artifact removal using independent component analysis (ICA). EEG epochs were extracted from 100 ms pre-stimulus onset (baseline) to 500 ms post-stimulus onset, and averaged epochs for each participant were baseline corrected. For further artifact reduction, we used the Multiple Artifact Rejection Algorithm (MARA; Winkler et al. 2011), an ICA-based tool, which allowed us to semi-automatically remove potential noise components.

Global Field Power Analysis

The global field power (GFP) serves as metric for the reference-independent strength of a neuronal response and is calculated by determining the standard deviation of potentials across all electrodes (Lehmann and Skrandies 1980). For our study, we computed the GFP individually for each participant and each condition. Subsequent analyses involved repeated-measures ANOVAs at each GFP time point, using group (high and low dominance) and condition (Happy vs. Sad,

Angry vs. Neutral) as the factors. Notably, we only regarded effects as reliable if they exhibited continuous significance over a span of 10 ms (Guthrie and Buchwald 1991).

Distributed Electrical Source Imaging

Source analysis was conducted by using the Fieldtrip software (Oostenveld et al., 2011). As we did not have fMRI scans for our experiment, we utilized the head and source models provided directly by Fieldtrip. The layout of the electrodes was aligned to the head model to generate a 3D electrode map. Subsequently, the EEG data was reconstructed using the linearly constrained minimum variance (LCMV) beamformer method, as integrated within the Fieldtrip package. Our primary objective was to pinpoint the disparities in source-space activity between the high and low dominance groups, particularly during the instances where the GFP peaks were observed. Hence, we employed two statistical analyses: a 2x2 ANOVA to assess the main effects of experimental conditions and dominance groups, as well as their interaction. Secondly, an independent t-test was applied to the reconstructed EEG data at each voxel, contrasting the activity levels between the two dominance groups. To correct for multiple comparisons, we used a spatial criterion requiring clusters to contain at least 15 neighboring solution points with significant effects ($p < 0.05$) (da Cruz et al., 2018).

Frequency and network analysis

In our resting state EEG data analysis, we employed Fieldtrip (Oostenveld et al., 2011) to perform frequency analysis utilizing Welch's method. This approach was applied across the entire 5-minute EEG dataset. The time series was segmented into 4-second windows, with each segment overlapping its neighboring window by 50%. Within each window, a fast Fourier transform was executed using a DPSS (discrete prolate spheroidal sequences) taper. By averaging over these windows, we obtained a comprehensive spectral estimate. The spectral data was then organized into four distinct frequency bands – delta (1-4Hz), theta (4-8Hz), alpha (8-12Hz), and beta (12-30Hz). The power was quantified for each band and subsequently averaged across all electrodes.

Further, we delved into the functional connectivity in the EEG data within the frequency domain. Leveraging Fieldtrip, we calculated coherence among electrode pairs for each frequency band. To mitigate potential artifacts from volume conduction, we considered only the imaginary part of coherence. The resulting coherence values gave rise to an assessment of network density, which served as a holistic representation of the overall connectivity in the EEG dataset.

For our statistical approach, we conducted a 4x2 ANOVA on both the power and network density values. This was aimed to identify the main effects associated with frequency bands and dominance groups, as well as any interaction effects between these two factors.

Mediation analysis

We first divided the brain topography into four lobes, frontal, parietal, temporal, and occipital (Bian et al., 2014). We then averaged the neural activities within each region and computed correlations between these activities and the PRF-D scores as well as reaction times, at each time points, to identify regions and times of interest.

Next, we conducted mediation analysis using PRF-D scores as the independent variable, reaction time as the dependent variable, and the difference between frontal and occipital activities as the mediating factor. Statistical analysis was performed using the Python toolbox 'penguin' to determine direct, indirect, and total effects, employing a bootstrapping method with 1000 iterations (Vallat, 2018).

Behavioral Experiments

A simple reaction time and a Go/NoGo task were performed by the participants (Fig. 6.4.2). In the simple reaction time task, a grey square appeared either to the left or right of the central cross. The inter-trial interval varied randomly between 0.15 and 1.5 seconds. Participants were instructed to press the spacebar as soon as the square appeared, irrespective of its position. If no response was given within 0.5 seconds, the square disappeared. A practice session of ten trials was given before the experiment proper, comprising 200 stimuli. To avoid fatigue, short breaks were introduced after every 50 trials. The task typically was performed in about 10 minutes.

In the Go/NoGo task, participants were presented with either a square or a triangle, serving distinctively as the Go or NoGo stimulus. Upon the display of the Go stimulus, participants were asked to hit the spacebar within 0.5 seconds. Conversely, they were to abstain from any response when confronted with the NoGo stimulus. Maintaining a 70:30 ratio, the Go stimulus was shown in 70% of the trials. Following the same procedure as the simple reaction time task, after a practice block of ten trials, a block of 200 stimuli was presented. Breaks were introduced every 50 trials, and the entire task took approximately 10 minutes.

To prevent any sequence bias, the order of the two tasks was pseudo-randomized, with half of the participants starting with the simple reaction time task and the other half commencing with the Go/NoGo task.

Salivary Cortisol Analyses

Participants provided three saliva samples during the study. The first immediately after they signed the consent form, the second at the beginning of condition 2, and the third approximately 20 minutes after the end of the second condition. The time interval between each collection was approximately 1 hour. Each collection gathered approximately 0.8 to 1.4 mL of saliva into 10 mL polypropylene tubes, which were then stored in a freezing environment below -20 ℃ until processing. Subsequently, these samples were centrifuged at a speed of 3000 rpm for a duration of 15 minutes under room temperature conditions. The resulting salivary cortisol and testosterone concentrations were measured using an enzyme immunoassay, following the guidelines provided by the manufacturer, Salimetrics. To account for cortisol's circadian rhythm, we ensured all experimental sessions took place between 1 PM and 7PM.

Cortisol and testosterone reactivity was evaluated by computing the indices for the area under the curve with respect to ground (AUCg) and the area under the curve with respect to increase (AUCi), following the methodology outlined by Pruessner et al. (2003). Given that the time intervals between each saliva collection were identical, we used the following formula to calculate AUCg:

$$AUC_g = \sum_{i=1}^{n-1} \frac{(m_{(i+1)} + m_i)}{2}$$

where n is the total number of measurements (in our case, n = 3), and m_i is the individual measurement of cortisol or testosterone. We calculated AUCi using the formula:

$$AUC_i = AUC_g - (n-1) * m_1$$

with the same notation as for AUCg.

Mixed-effect models

To understand the influence of various factors on the participants' behavioral performance, we applied mixed-effect models to incorporate potential variables with JASP. In Model 1 for reaction times, we included accuracy, trait anxiety, experimental conditions, and dominance groups as fixed effects. Model 2 was extended to incorporate an interaction term between experimental conditions and dominance groups.

6.4 *Results*

Participants with scores greater than or equal to the median were defined as the high dominance group, and those with scores less than the median are assigned to the low dominance group. First, we did not find that dominant females are faster than less dominant ones (Fig. 6.4.1B, $F(1,22.36)=2.47$, $p=0.13$, partial $\eta^2 = 0.1$), contrary to the male population in da Cruz et al. (2018). Despite lacking statistical significance, the direction of the differences aligns with observations within the male population. This trend is further shown in the correlation between the level of dominance and reaction time, especially in the Happy-Sad condition (Fig. 6.4.1D, $r(24) = -0.36$, $p = 0.068$).

In addition to reaction time, we also utilized the mixed-effect model to analyze accuracy. A significant main effect was observed for the condition on accuracy. Specifically, the participants demonstrated higher accuracy in the Happy-Sad condition when compared to the Angry-Neutral condition (Fig. 6.4.1C, $F(1,22)=48.28$, $p<0.01$, partial $\eta^2 = 0.66$). However, we observed no main effect of dominance group on accuracy (Fig. 6.4.1C, $F(1,21.35)=1.59$, $p=0.22$, partial $\eta^2 = 0.069$).

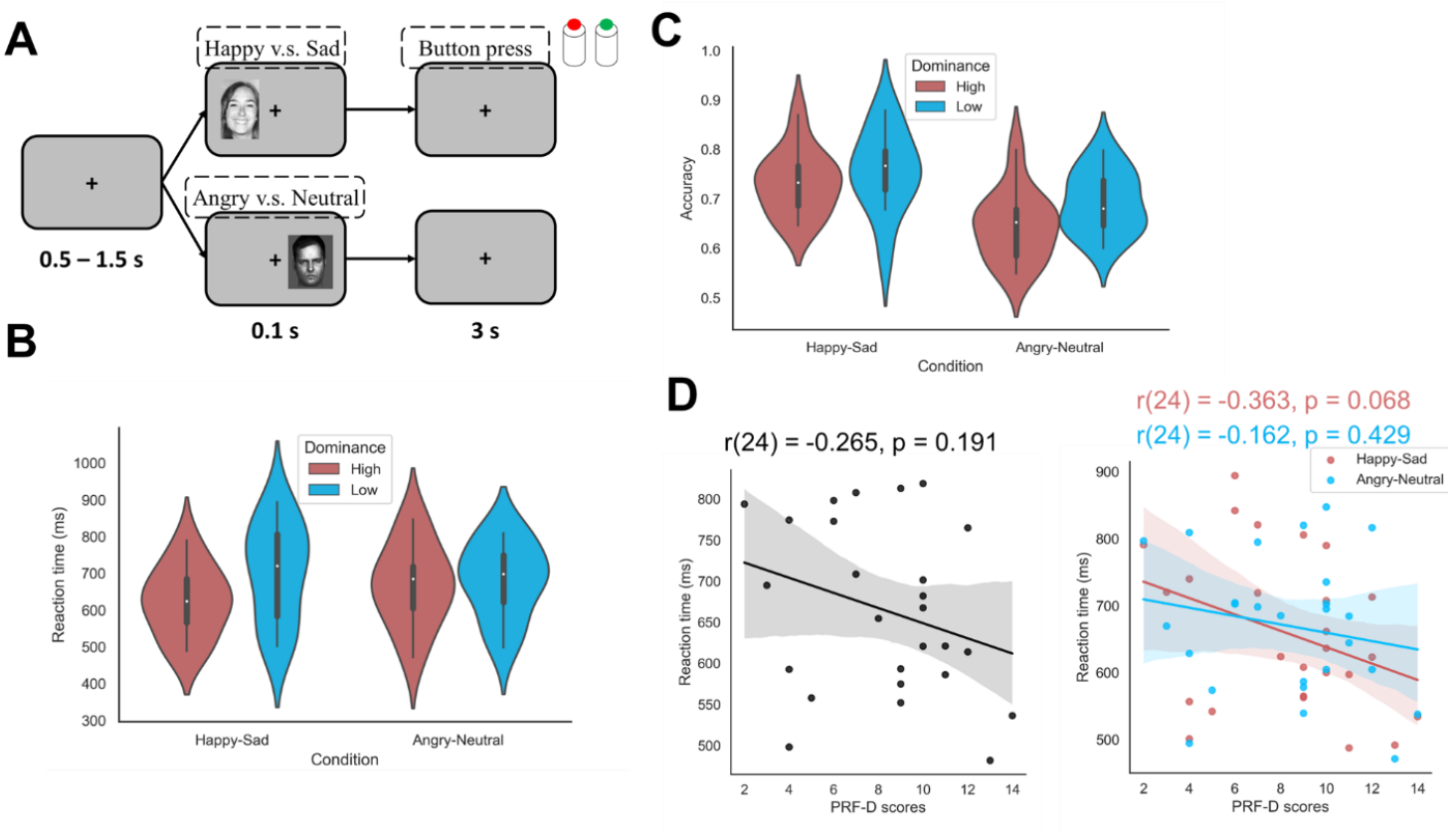

Figure 6.4.1. Behavioral results in the emotion discrimination task. (A) The experimental design of the task. There are two experimental conditions: Happy-Sad and Angry-Neutral. Participants are required to press one of the two buttons corresponding to the emotion they perceived as soon and as accurate as possible. (B) Violin plots showing the reaction times of both the high and low dominant group in the two experimental conditions. (C) Violin plot showing the accuracy of both the high and low dominant group in the two experimental condition. (D) Scatter plot displaying the correlation between PRF-D scores and reaction times.

Nevertheless, following the same approach as da Cruz et al. (2018), we employed a simple reaction time task to examine if motor ability differences existed between the dominance groups (Fig. 6.4.2A). Although a trend suggested that individuals with high dominance responded slower than those with lower dominance, no significant differences emerged in the reaction times between the two dominance groups, suggesting comparable motor skills (Fig. 6.4.2B, $t(24) = 1.31$, $p = 0.2$, Cohen's $d = 0.52$). We further examined any potential differences in impulsivity between the groups by conducting a Go/No-Go task (Fig. 6.4.2C). We observed no significant differences in the reaction times for hit trials ($t(24) = 0.87$, $p = 0.4$, Cohen's $d = 0.34$) or in the d-prime score ($t(24) = -0.42$, $p = 0.68$, Cohen's $d = 0.17$) between the groups (Fig. 6.4.2D). These findings

indicate that superior motor abilities or increased impulsivity are not characteristics that discriminate the high dominance group.

The relationship between hormone levels and dominance has been reported previously. For instance, Mehta and Josephs (2010) demonstrated that testosterone is positively correlated with dominance, but only in individuals with low levels of cortisol. Therefore, we evaluated if relevant hormonal differences existed between the two dominance groups. Specifically, we calculated the area under the curve with respect to the ground (AUC_g) and the increase (AUC_i) for two hormones, cortisol and testosterone (Fig. 6.4.S1). There were no significant differences between the two groups.

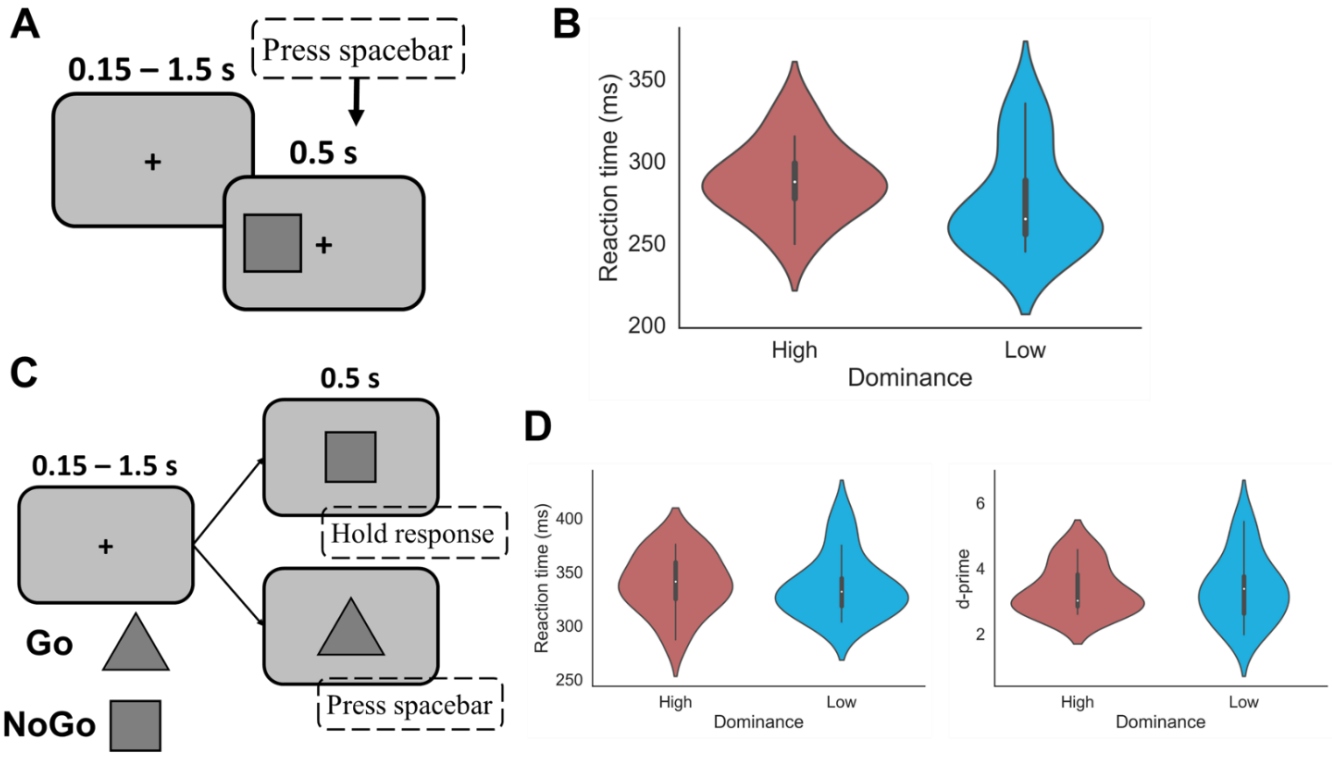

Figure 6.4.2. Results of simple reaction time and Go/NoGo tasks. (A) The experimental design of the simple reaction time task. Participants pressed the spacebar as soon as they noticed the onset of a square on the screen. (B) Violin plots showing the reaction times of both the high and low dominant groups. (C) The experimental design of the Go/No-Go task. A Go or a No-Go stimuli appeared on the screen. Participants were required to press the spacebar when a Go stimulus was presented and withhold their response when the No-Go stimulus was presented (D) Left: Violin plot showing the reaction times of hit trials both high and low dominant groups. Right: Violin plot showing the d-prime of both high and low dominant groups.

EEG analysis. To understand the neural mechanism underlying social dominance, we recorded EEG signals while participants performed the emotion discrimination task. We found differences in the GFP between the two dominance groups. These differences emerged within a time interval, ranging from 238ms to 251ms after the stimulus presentation (Fig. 6.4.3A). Our findings align with the results reported previously in the male population (da Cruz et al., 2018).

To better understand which electrodes contribute to the difference observed in the GFP, we conducted a repeated measures ANOVA on the neural activity averaged across the previously reported period (238-251ms) for each electrode (128 electrodes), with dominance level and experimental conditions as factors. A main effect of the dominance factor was observed in six electrodes, while an interaction between dominance level and experimental conditions was evident in twenty electrodes. (Fig. 6.4.3B, top). Although the previously mentioned effects (six for the main effect of dominance and twenty for interaction effect) did not survive correction for multiple comparisons, several demonstrated large effect sizes. This suggests that these regions may encode different levels of dominance and could contribute to the differences in the GFP we observed.

We also carried out an exploratory analysis aimed to determine if there were any EEG components that linked dominance levels to behavior, particularly reaction time. We observed significant correlations between both PRF-D scores ($r(24) = -0.445$, $p = 0.023$) and reaction times ($r(24) = 0.57$, $p = 0.002$) with the differences in neural activities between frontal and occipital lobes from 62ms to 76ms after stimulus presentation (Fig. 6.4.S2B). Subsequently, we applied a mediation analysis and observed a significant indirect effect along the path, with PRF-D as the independent variable, reaction time as the dependent variable, and the difference between the two brain regions as the mediation factor. This suggests that the dominance level is not a direct cause of reaction time, but is mediated by the neural activities, which in turn affect the response speed (Fig. 6.4.S1C).

Furthermore, source level analysis localizes the differences in the left intraparietal sulcus (IPS) (Fig. 6.4.3B). To capture the differences at the neural level more effectively, we compared the high and low dominance groups separately across the two experimental conditions. In the

Happy-Sad condition, differences were linked to regions in the striatum. Conversely, in the Angry-Neutral condition, differences were associated with the temporal regions (Fig. 6.4.S3).

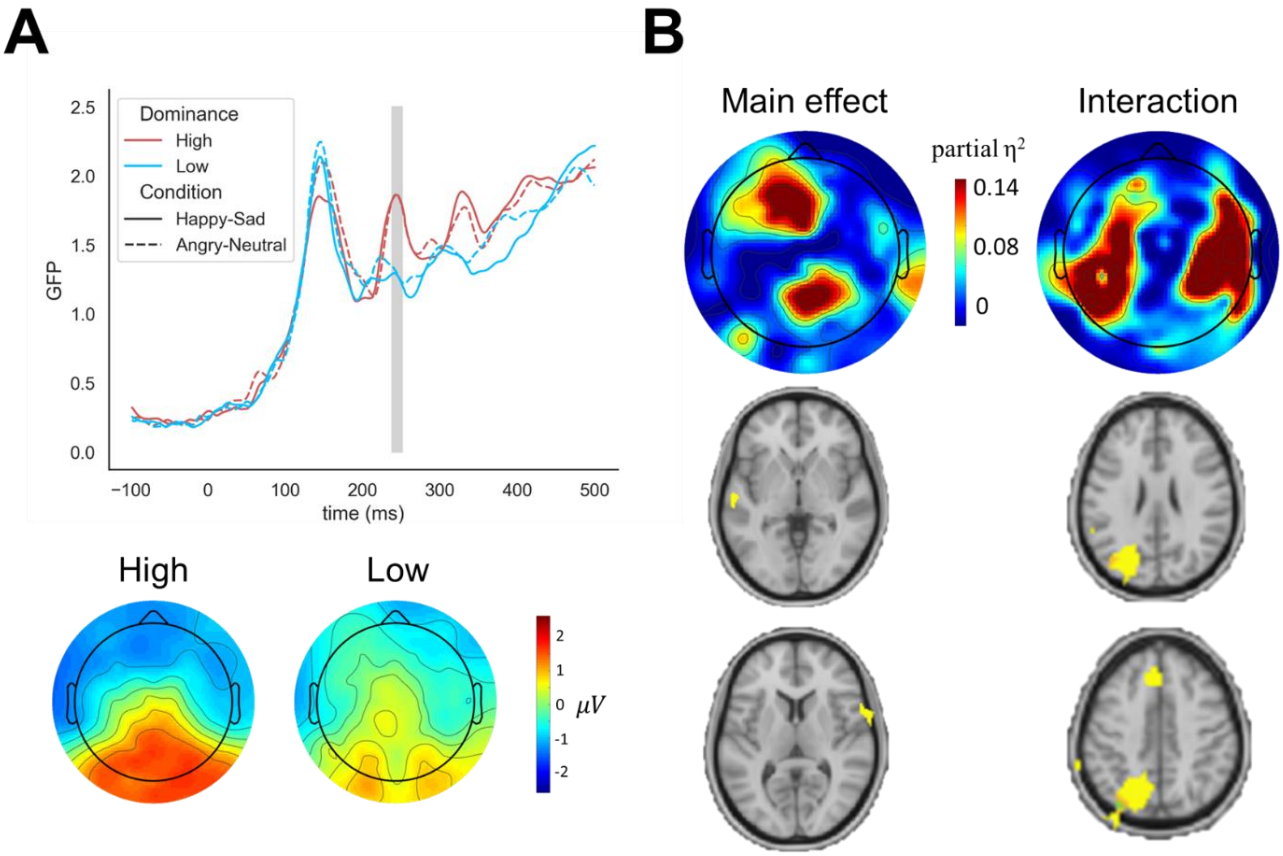

Figure 6.4.3. High dominant females exhibit a distinctive EEG component. (A) Top: The GFP traces for the two groups in the two experimental conditions. The gray region marks the time points that show the main effect of Group. Bottom: The averaged topographies of the two groups. (B) Left: The main effect of the groups at topographic and source levels. Right: The interaction effect of groups and condition at the topographic and source level.

After the main experiment, we administered a 5-minute session of resting state EEG recording. While we observed no significant differences in the power across frequency bands in the two groups (Fig. 6.4.4A, $F(1,96) = 1.51$, $p = 0.22$, partial $\eta^2 = 0.015$), the whole-brain network density, which was computed from the connectivity analysis, revealed a main effect associated with dominance groups (Fig. 6.4.4B, $F(1,96) = 5.17$, $p = 0.025$, partial $\eta^2 = 0.051$). These findings suggested that the high dominance group exhibited a greater network density relative to the low dominance group.

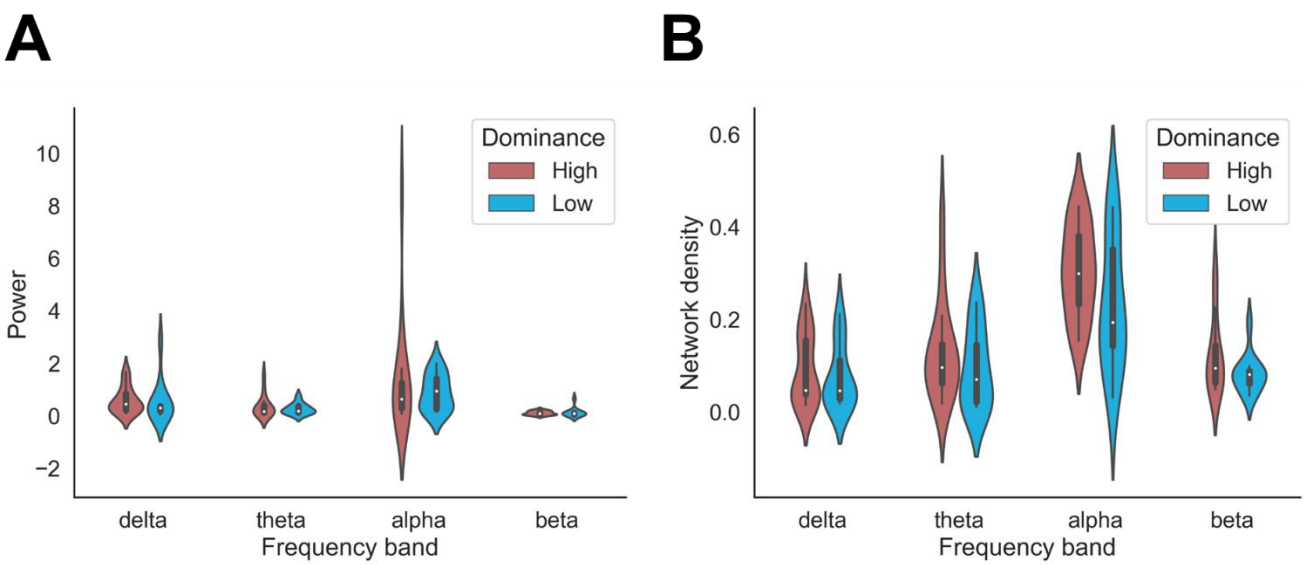

Figure 6.4.4. Social dominance and resting state EEG. (A) The power of frequency bands on the two groups in four frequency bands. (B) The network density in the two groups for four frequency bands.

6.5 *Discussion*

Previous work showed that dominant males respond faster in complex decision-making paradigms compared to subordinate ones, but such differences do not hold in simpler reaction time paradigms (da Cruz et al., 2018). During these complex tasks, the EEG N2/P2 component was notably stronger in high dominant males than in less dominant ones. In the current study, we tested whether similar patterns occur in female participants. The amplitude and peak time of the N2/P2 component in dominant females closely resembles those in dominant males. However, we did not find a significant reaction time benefit for high dominance vs. low dominant females. These results further support again the view that the N2/P2 component is a trait marker for dominance, evident even in the absence of social interactions.

We additionally observed that activity around 70ms after stimulus presentation seemed to be important for the linkage between dominance levels and behavior, especially involving the frontal and occipital lobe. It was reported that the N70 is more pronounced in females than in males (Breton et al., 2019). In addition, the N70 also seems to relate to the processing of emotional faces (Santos et al., 2008; Zotto and Pegna, 2015). These studies relate to our findings that the EEG component around 70ms after the presentation of an emotional facial stimulus might be associated with processing speed of emotional stimuli, particularly in female participants.

While we did not observe a statistically significant difference in reaction times, there was a noticeable correlation between dominance scores and reaction time. This suggests a trend for dominant females to respond more quickly. Some studies suggest that social dominance may be positively linked to impulsivity (Lesch and Merschdorf, 2000) To address the possibility and to control for the motor ability, participants performed two additional behavioral tasks, the simple reaction time and the Go/NoGo tasks. There were no significant differences between the two groups in these tasks.

Behaviorally, we did not observe faster reaction times in dominant females, unlike the observation from da Cruz et al., (2018) with males. This discrepancy might be attributed to differences in the sampling methods applied in the studies. da Cruz et al., (2018) used an extreme sampling approach, selecting participants from the two ends of the dominance spectrum (very high and very low). In contrast, our study utilized a random sampling method, resulting in a normally

distributed range of dominance levels among participants. This difference in sampling strategy could contribute to the smaller effect size observed in our behavioral results.

On the other hand, in the female population, we observed an N2/P2 EEG component that is almost identical to the study by da Cruz et al., (2018) in both amplitude and latency. There are studies showing that the N2/P2 component is related to attention (Schindler and Bublatzky, 2020). Therefore, it may be that dominant individuals have higher attention resources. We performed source localization to identify the brain regions that contributed to the observed differences in the N2/P2 EEG component between the two dominance groups. Our findings indicate that the IPS plays a significant role in these differences. The IPS, a region previously implicated in the encoding of social status, exhibits stronger activity when recognizing stimuli associated with a higher social hierarchy (Chiao et al., 2009; Farrow et al., 2011). This finding may be reflected in our study, suggesting that high dominant females might allocate more attention resources to social judgments during our task. When contrasting the neural activities under the two experimental conditions separately, we observed that regions in striatum (Happy-Sad condition) and temporal sulcus (Angry-Neutral) exhibited differences between the high and low dominant groups. These brain regions were also reported in the study by da Cruz et al. (2018). This result suggests that different process might be occurring in the two conditions. The temporal sulcus, crucial for encoding facial information (Å hs et al., 2014), showed higher activity in high dominant females, which could be interpreted as indicative of increased attention resources used for discriminating faces. On the other hand, the greater response in the striatum might be related to heightened perception of social rewards (Chakrabarti et al., 2006), potentially linked to the presence of happy faces, eliciting a stronger reaction in high dominant females.

The resting-state EEG is an effective measurement for assessing fundamental neural processing in the brain. For this purpose, we analyzed resting-state EEG data to determine whether significant differences existed between the two levels of dominance in the female participants. Focusing on the connectivity between electrodes, we observed a significant difference in connectivity density – the high dominant group demonstrated denser connectivity compared to the low dominant group. While there have been studies utilizing resting-state EEG to examine trait aggressive tendencies (Hofman and Schutter, 2012) and combining the lateralized resting activity with questionnaire (BIS/BAS) to make prediction of neural responses when view emotional faces

(Balconi et al., 2017), no previous studies have directly investigated the relationship between social dominance and resting-state brain connectivity. Thus, our study provides a new direction by showing that inherent differences in brain connectivity can be indicative of an individual's level of social dominance, independent of task engagement.

In conclusion, our study provides evidence that dominant females have a higher N2/P2 EEG component compared to the low dominant females during decision-making. This mirrors the findings of da Cruz et al. (2018), where the same component was presented in males. This gives a strong evidence of a general neuromarker for social dominance. Social dominance is often associated with leadership and decision-making ability. Our results provide evidence that such tendencies might be innate, suggesting that some individuals could be naturally predisposed to leadership. Moreover, these tendencies are not confined to demographic factors like gender, rather, they appear to be a general property.

Supplementary results

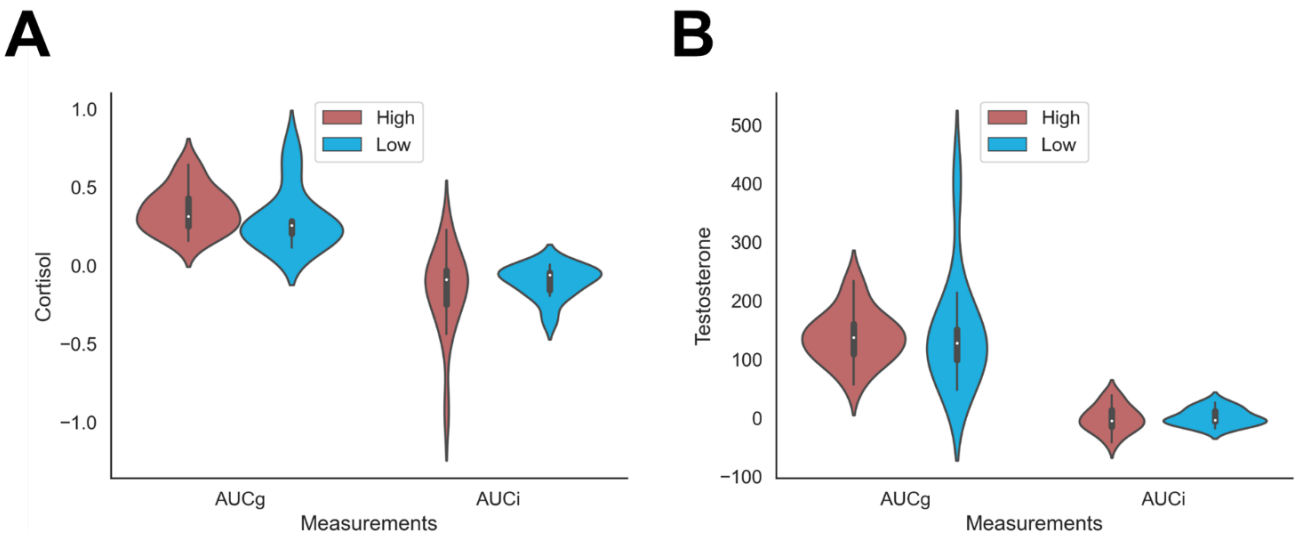

Figure 6.4.S1. Hormonal levels. (A) Violin plot showing the AUC_g and AUC_i of cortisol of both high and low dominant groups. (B) Violin plot showing the AUC_g and AUC_i of testosterone of both high and low dominant groups. Further details can be found in the Method: Salivary Cortisol Analyses' section.

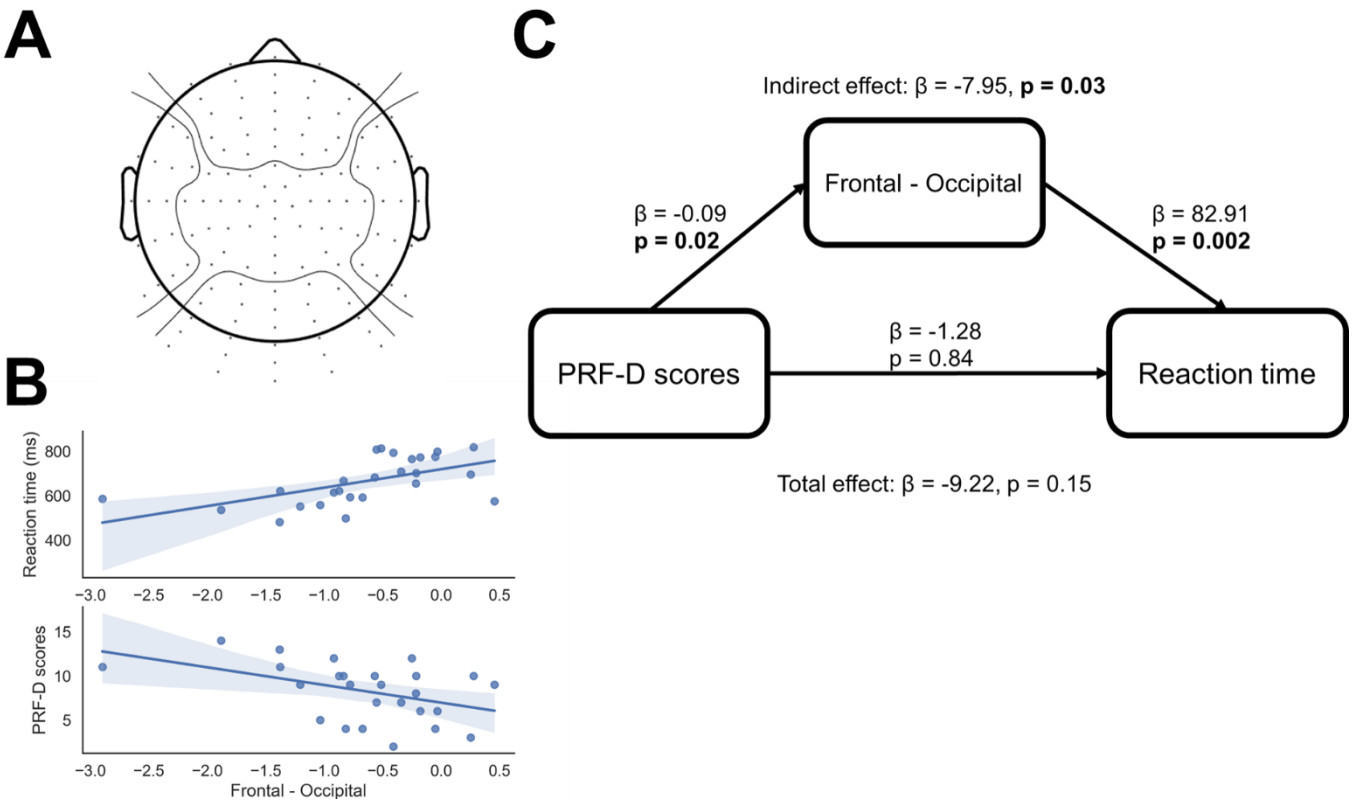

Figure 6.4.S2. Mediation analysis. (A) Division of EEG topography into four regions. (B) Correlation between the differences in average activities of the frontal and occipital lobes with the PRF-D scores and reaction times. (C) Diagram illustrating the mediation analysis.

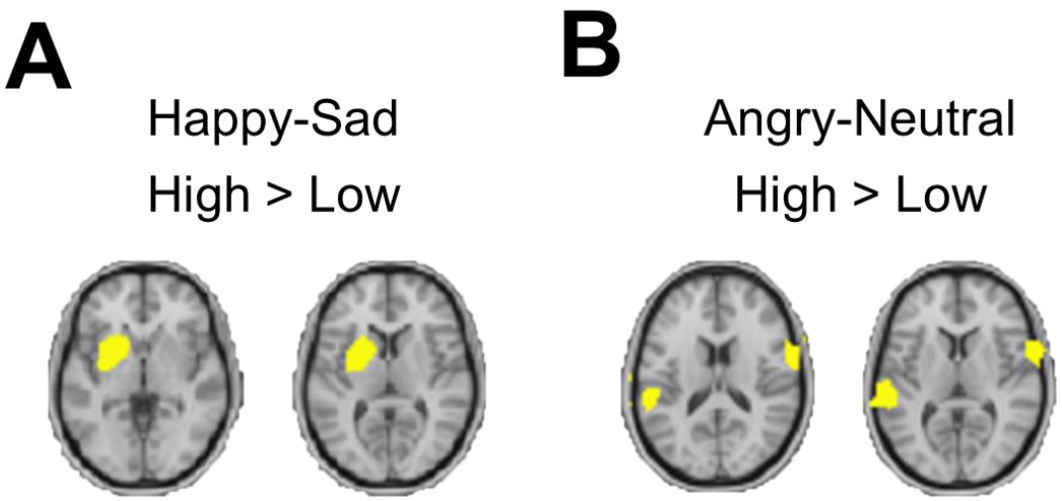

Figure 6.4.S3. Condition-specific source analysis. (A) The contrast of source between the high and low dominance groups during the Happy-Sad condition. (B) The contrast of source between the high and low dominance groups during the Angry-Neutral condition.

6.6 *References*

Å hs F, Engman J, Persson J, Larsson E-M, Wikström J, Kumlien E, Fredrikson M. 2014. Medial temporal lobe resection attenuates superior temporal sulcus response to faces. Neuropsychologia 61:291–298. doi:10.1016/j.neuropsychologia.2014.06.030

Anderson C, Kilduff GJ. 2009. Why do dominant personalities attain influence in face-to-face groups? The competence-signaling effects of trait dominance. J Pers Soc Psychol 96:491–503. doi:10.1037/a0014201

Anwyl-Irvine AL, Massonnié J, Flitton A, Kirkham N, Evershed JK. 2020. Gorilla in our midst: An online behavioral experiment builder. Behav Res Methods 52:388–407. doi:10.3758/s13428-019-01237-x

Balconi M, Vanutelli ME, Grippa E. 2017. Resting state and personality component (BIS/BAS) predict the brain activity (EEG and fNIRS measure) in response to emotional cues. Brain Behav 7:e00686. doi:10.1002/brb3.686

Bian Z, Li Q, Wang L, Lu C, Yin S, Li X. 2014. Relative power and coherence of EEG series are related to amnestic mild cognitive impairment in diabetes. Front Aging Neurosci 6. doi:10.3389/fnagi.2014.00011

Breton A, Ligneul R, Jerbi K, George N, Baudouin J-Y, Van Der Henst J-B. 2019. How occupational status influences the processing of faces: An EEG study. Neuropsychologia 122:125–135. doi:10.1016/j.neuropsychologia.2018.09.010

Chakrabarti B, Kent L, Suckling J, Bullmore E, Baron-Cohen S. 2006. Variations in the human cannabinoid receptor (CNR1) gene modulate striatal responses to happy faces. Eur J Neurosci 23:1944–1948. doi:10.1111/j.1460-9568.2006.04697.x

Chiao JY, Harada T, Oby ER, Li Z, Parrish T, Bridge DJ. 2009. Neural representations of social status hierarchy in human inferior parietal cortex. Neuropsychologia 47:354–363. doi:10.1016/j.neuropsychologia.2008.09.023

da Cruz J, Rodrigues J, Thoresen JC, Chicherov V, Figueiredo P, Herzog MH, Sandi C. 2018. Dominant men are faster in decision-making situations and exhibit a distinct neural signal for promptness. Cereb Cortex 28:3740–3751. doi:10.1093/cercor/bhy195

Da Cruz JR, Chicherov V, Herzog MH, Figueiredo P. 2018. An automatic pre-processing pipeline for EEG analysis (APP) based on robust statistics. Clin Neurophysiol 129:1427–1437. doi:10.1016/j.clinph.2018.04.600

Farrow TFD, Jones SC, Kaylor-Hughes CJ, Wilkinson ID, Woodruff PWR, Hunter MD, Spence SA. 2011. Higher or lower? The functional anatomy of perceived allocentric social hierarchies. NeuroImage 57:1552–1560. doi:10.1016/j.neuroimage.2011.05.069

Hall JA, Coats EJ, LeBeau LS. 2005. Nonverbal behavior and the vertical dimension of social relations: a meta-analysis. Psychol Bull 131:898.

Hofman D, Schutter DJLG. 2012. Asymmetrical frontal resting-state beta oscillations predict trait aggressive tendencies and behavioral inhibition. Soc Cogn Affect Neurosci 7:850–857. doi:10.1093/scan/nsr060

Jackson DN. 1974. Personality Research Form Manual. Research Psychologists Press.

Karamihalev S, Brivio E, Flachskamm C, Stoffel R, Schmidt MV, Chen A. 2020. Social dominance mediates behavioral adaptation to chronic stress in a sex-specific manner. eLife 9:e58723. doi:10.7554/eLife.58723

Lesch KP, Merschdorf U. 2000. Impulsivity, aggression, and serotonin: a molecular psychobiological perspective. Behav Sci Law 18:581–604. doi:10.1002/1099-0798(200010)18:5<581::AID-BSL411>3.0.CO;2-L

Ligneul R, Obeso I, Ruff CC, Dreher J-C. 2016. Dynamical Representation of Dominance Relationships in the Human Rostromedial Prefrontal Cortex. Curr Biol 26:3107–3115. doi:10.1016/j.cub.2016.09.015

Mast MS. 2002. Dominance as Expressed and Inferred Through Speaking Time.: A Meta-Analysis. Hum Commun Res 28:420–450. doi:10.1111/j.1468-2958.2002.tb00814.x

Mehta PH, Josephs RA. 2010. Testosterone and cortisol jointly regulate dominance: Evidence for a dual-hormone hypothesis. Horm Behav 58:898–906. doi:10.1016/j.yhbeh.2010.08.020

Mudar RA, Chiang H-S, Eroh J, Nguyen LT, Maguire MJ, Spence JS, Kung F, Kraut MA, Hart J. 2016. The Effects of Amnestic Mild Cognitive Impairment on Go/NoGo Semantic Categorization Task Performance and Event-Related Potentials. J Alzheimers Dis 50:577–590. doi:10.3233/JAD-150586

Oostenveld R, Fries P, Maris E, Schoffelen J-M. 2011. FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. Comput Intell Neurosci.

Pratto F, Sidanius J, Stallworth LM, Malle BF. 1994. Social dominance orientation: A personality variable predicting social and political attitudes. J Pers Soc Psychol 67:741–763. doi:10.1037/0022-3514.67.4.741

Pratto F, Stallworth LM, Sidanius J. 1997. The gender gap: Differences in political attitudes and social dominance orientation. Br J Soc Psychol 36:49–68.

Pruessner JC, Kirschbaum C, Meinlschmid G, Hellhammer DH. 2003. Two formulas for computation of the area under the curve represent measures of total hormone concentration versus time-dependent change. Psychoneuroendocrinology 28:916–931. doi:10.1016/S0306-4530(02)00108-7

Redhead D, Cheng JT, Driver C, Foulsham T, O'Gorman R. 2019. On the dynamics of social hierarchy: A longitudinal investigation of the rise and fall of prestige, dominance, and social rank in naturalistic task groups. Evol Hum Behav 40:222–234. doi:10.1016/j.evolhumbehav.2018.12.001

Santamaría-García H, Pannunzi M, Ayneto A, Deco G, Sebastián-Gallés N. 2014. 'If you are good, I get better': the role of social hierarchy in perceptual decision-making. Soc Cogn Affect Neurosci 9:1489–1497. doi:10.1093/scan/nst133

Santos IM, Iglesias J, Olivares EI, Young AW. 2008. Differential effects of object-based attention on evoked potentials to fearful and disgusted faces. Neuropsychologia 46:1468–1479. doi:10.1016/j.neuropsychologia.2007.12.024

Scheggia D, La Greca F, Maltese F, Chiacchierini G, Italia M, Molent C, Bernardi F, Coccia G, Carrano N, Zianni E, Gardoni F, Di Luca M, Papaleo F. 2022. Reciprocal cortico-amygdala connections regulate prosocial and selfish choices in mice. Nat Neurosci 25:1505–1518. doi:10.1038/s41593-022-01179-2

Schindler S, Bublatzky F. 2020. Attention and emotion: An integrative review of emotional face processing as a function of attention. Cortex 130:362–386. doi:10.1016/j.cortex.2020.06.010

Schutz WC. 1958. FIRO: A three-dimensional theory of interpersonal behavior.

Spielberger CD. 1983. State-trait anxiety inventory for adults.

Vallat R. 2018. Pingouin: statistics in Python. J Open Source Softw 3:1026. doi:10.21105/joss.01026

Winkler I, Haufe S, Tangermann M. 2011. Automatic classification of artifactual ICA-components for artifact removal in EEG signals. Behav Brain Funct 7:1–15.

Zhou T, Sandi C, Hu H. 2018. Advances in understanding neural mechanisms of social dominance. Curr Opin Neurobiol 49:99–107. doi:10.1016/j.conb.2018.01.006

Zink CF, Tong Y, Chen Q, Bassett DS, Stein JL, Meyer-Lindenberg A. 2008. Know Your Place: Neural Processing of Social Hierarchy in Humans. Neuron 58:273–283. doi:10.1016/j.neuron.2008.01.025

Zotto MD, Pegna AJ. 2015. Processing of masked and unmasked emotional faces under different attentional conditions: an electrophysiological investigation. Front Psychol 6. doi:10.3389/fpsyg.2015.01691

7. Discussion

In our fast-changing environment, possessing the abilities to learn and make decisions is crucial. In the thesis, we provide evidence from behavioral and neural perspectives to understand how various factors affect these capabilities.

Novelty seeking is important for reinforcement learning

In study I, we introduced a unique experimental paradigm to investigate human exploration driven by curiosity under stochastic conditions. Two key findings arose from this study: 1) Participants who exhibited optimism about encountering greater rewards than those already found consistently displayed distraction due to stochastic elements; and 2) this continuous distraction pattern was more effectively attributed to the novelty seeking rather than the pursuit of information-gain or surprise, despite the theoretical robustness of information-gain in managing stochastic scenarios.

Reward is a crucial component in various forms of learning (Berridge & Robinson, 1998). The expectation of rewards hidden in an environment often evokes optimism in humans, as it leads to more positive expectations about the future outcomes (Carver et al., 2010). In our experiment, we implicitly manipulated participants' levels of optimism by varying the magnitude of the rewards they encountered. Indeed, our results demonstrated that participants with higher optimism (holding the expectation of finding better reward) tended to explore more in stochastic states, likely influenced by an optimism bias (Sharot, 2011).

In the realm of reinforcement learning, the exploration-exploitation trade-off is a crucial concept extensively explored across psychology, neuroscience, and machine learning (Cohen et al., 2007; Dubey & Griffiths, 2020). Numerous theories and models have been developed to understand how humans formulate strategies to explore the environment (Ghavamzadeh et al., 2016; Schulz & Gershman, 2019). In our studies, we specifically evaluated three prominent models centered on novelty, surprise, and information-gain. We discovered that high levels of exploration in stochastic states are most accurately explained by the novelty-seeking model. This finding implies that humans might employ a less complex algorithm in such scenarios, as they only need to track state frequencies over time without requiring knowledge of the environmental structure.

Therefore, while this approach may not always be optimally efficient, it is less demanding cognitively (Xu et al., 2021).

The link between dopamine and reinforcement learning

As highlighted in the introduction, dopamine is one of the most crucial neurotransmitters related to RL (Schultz et al., 1997). Accordingly, our subsequent studies aimed to investigate how potential deficit in the dopamine system might impact an individual's RL ability. Our primary focus was on two populations: the elderly and patients with schizophrenia.

In study II, we introduced a model based on classical Q-learning, modified to simulate elevated dopamine levels in schizophrenia. We proposed three hypotheses: 1) Patients with schizophrenia would reveal suboptimal RL performance compared to healthy controls when faced with positive rewards; 2) This suboptimal performance would stem from poor action selection; and 3) These patients would exhibit minimal or no deficits in RL tasks involving negative rewards, i.e., punishments.

Our results indicate that the suboptimal accuracy in patients could be due to high perseverative behavior. Within the Q-learning framework, such perseveration is linked to a reduced encoding of prediction errors, leading to a diminished magnitude of updates in the values associated with state-action pairs and, consequently, slower learning. In the model, the inverse temperature parameter (indicating exploration rate) is higher in patients than in controls. Thus, unlike controls, patients tend to meander through the environment until they fortuitously achieved the goal. We suggest that the model effectively reflects the underlying mechanisms in these patients. Our results are consistent with the studies demonstrating that patients exhibit deficits in reward-based learning (Strauss et al., 2011, 2014).

In experiment 2 of the same study, patients exhibit significantly higher accuracy in the punishment condition than in the reward condition. These findings align with previous research, which reported impairments in RL with positive rewards but not with punishments (Abohamza et al., 2020; G. L. F. Cheng et al., 2012; Waltz et al., 2007b). Another study observed deficits in forming new stimulus-reward associations, whereas patients performed normally in tasks involving new stimulus-punishment associations (Waltz et al., 2007b).

In study III, our objective was to explore potential deficits in RL among older adults. In the first experiment, no significant differences in RL abilities are observed between young and older adults. Even when increasing the number of trials, as in experiment 2, we merely observe a difference in one of the ISI conditions These results suggest that, generally, older adults do not exhibit inferior RL abilities compared to young adults, aligning with the findings from previous studies (Lighthall et al., 2018; Pietschmann et al., 2011).

However, other research has shown that older adults may demonstrate reduced RL in more demanding tasks (Daniel et al., 2020). This conclusion might be influenced by a ceiling effect in a simpler condition. It is possible that group effects are present in both the easier and more challenging conditions, suggesting that the difficulty does not appear to be a determining factor. Our findings imply that RL abilities are largely preserved in older adults within our experimental framework. Importantly, our study underscores the significance of reporting null results in aging research, as they are crucial in accurately understanding the impact of aging on human cognition.

Different effects of aging and schizophrenia on the ability to engage in RL

In study II and III, we utilized the same RL task to two distinct populations: patients with schizophrenia and older adults. Our analysis enabled us to distinguish and observe differences in their behavior. Among the patient population, we observed a higher tendency for perseveration compared to the healthy controls, which appear to result from dysfunctions in the dopamine system. One could argue that such perseveration might also arise from reduced working memory capacity, leading to confusion about which actions are appropriate. To address this, we examined performance under two different ISI conditions within the experiment. We observed no interaction effects indicating that patients and controls performed differently in the two ISI conditions. Given that the long ISI condition demands greater memory capacity, the absence of any significant differences suggests that the suboptimal performance is not attribute to deficiencies in the memory system.

In the study III, we found significant performance differences between young and older adults in the long ISI condition. This could be due to several factors: actual RL deficits, memory issues, or greater fatigue in older adults. First, the Q-learning model showed no abnormal learning or exploration rates in older participants, suggesting that RL deficits are unlikely. Second, older

adults showed less confidence in memory tasks involving distant states and had less improvement in decision variability, hinting at possible memory challenges. Similar studies, like Lighthall et al. (2018), have found differences in brain activity but not in behavior between age groups, suggesting that memory load might affect the brain but not always be evident in behavior. Third, it's known that older people tire faster (Enoka & Duchateau, 2016), which could explain the differences in long ISI tasks. However, our analysis showed that older adults struggled more with tasks involving distant goals but not with closer ones. This pattern points more towards a memory deficit being the main issue rather than fatigue.

In sum, we identified deficits in RL in both populations, though the strength and the causes of these deficits varied. Patients with schizophrenia exhibited more pronounced deficits, likely resulting from increased perseveration due to abnormally high dopamine levels. On the other hand, the ageing population showed subtler differences, potentially arising from issues within the memory system.

After exploring how novelty, illness, and aging influence RL and decision-making, we now shift our focus to a more intrinsic factor: personality traits. Our aim is to determine whether and how these traits impact behavior.

How social dominance, as a trait, affects behavior

Humans inhabit a society that is highly interconnected. Within this context, hierarchical structures and the concept of social dominance emerge (Sidanius & Pratto, 1999). These concepts, crucial in defining the ranks among individuals, significantly influence human behavior, including learning and decision making (Hall et al., 2005; Ligneul et al., 2017; Zink et al., 2008). Nevertheless, some research has shown that social dominance can be considered as a trait, independent of social context. For instance, dominant male participants have demonstrated faster responses in complex decision-making paradigms compared to simpler reaction time tasks (da Cruz et al., 2018). In these complex tasks, the EEG N2/P2 component was more pronounced in highly dominant males than in less dominant ones. Building on this, study IV aimed to explore whether female participants exhibit similar patterns, both behaviorally and neurologically.

We utilized the same emotion discrimination task with EEG recordings as in da Cruz et al. (2018) to test the female participants. The EEG results showed that the amplitude and peak latency

of the N2/P2 component in dominant females closely resembles those in dominant males. However, we did not observe a reaction time advantage for high dominance compared to low dominance in females. Additionally, participants undertook two other behavioral tasks: the simple reaction time and the Go/NoGo tasks. No significant differences were found between the two dominance groups in these tasks. We also examined the resting-state EEG recordings. Dominant females exhibited increased network density across all frequency bands.

The N2/P2 component observed in EEG recordings is thought to represent the allocation of cognitive resources towards the detection, categorization of target stimuli, and associated decision-making processes (Mudar et al. 2016). It may also reflect the intentional effort required to execute the task (Winterer et al. 2002). Consequently, variations in N2/P2 activation related to dominance could indicate differences in the mobilization of these cognitive resources. The intrinsic disparities in brain processes might underpin distinctive cognitive styles that contribute to the establishment of dominance hierarchies and leader-follower dynamics in social settings.

The results of our study are in line with those found by da Cruz et al. (2018). We observed an increased N2/P2 EEG component in dominant females, suggesting that this component could serve as a general neuromarker for social dominance, even in the absence of a social context. Furthermore, our findings revealed increased whole-brain connectivity in dominant females during resting state. This observation suggests fundamental neural wiring differences between individuals of high and low dominance even without task engagement.

General discussion

The results of this thesis provide several insights into the fields of RL and decision-making. Firstly, we analyzed the source of reduced RL performance, distinguishing between the effects of dopamine and memory. This was achieved by studying both patients with schizophrenia and a healthy aging population. Previous studies such as Strauss et al. (2011) and Dowd et al. (2016) have found that patients show less accuracy in RL tasks involving rewards. Furthermore, studies by Waltz et al. (2007a) and Cheng et al. (2012) demonstrated better performance by these patients in RL tasks involving punishment. Our findings not only corroborate the observed deficits in reward learning but also suggest intact performance in punishment learning. The discovery of unimpaired performance in punishment learning is astonishing and strengthens the proposition of

our model. It provides a clear explanation by applying Q-learning model to illustrate the impact of abnormal dopamine concentrations. While some studies have suggested that the ageing population demonstrates less accuracy in RL due to issues in dopamine (Chowdhury et al., 2013; Sojitra et al., 2018), the results are not consistent. Lighthall et al. (2018) and Eppinger et al. (2008), for instance, found no significant or only subtle differences in RL accuracy between older and young individuals. Our results align more with the latter, as we observed only minor differences in performance between the young and older adults. We determined that the RL deficits in schizophrenia are linked to how dopamine encodes information, while the ageing population is primarily affected by a reduced memory capacity, which influences their ability to choose the correct actions.

Another key finding of our studies concerns intrinsic motivation and its influence on our interactions with environment. Previous studies have demonstrated that humans are naturally drawn to novelty or surprising elements, which enhances exploratory behavior in the RL framework (Houillon et al., 2013; Itti & Baldi, 2009; Wittmann et al., 2008). Furthermore, humans generally exhibit optimism, often phrased as optimistic bias, when expecting better outcomes (Lefebvre et al., 2017). Our research revealed that, despite being suboptimal, humans tend to seek novelty, and the expected rewards reinforces such behavior. This tendency is also captured in our RL model. These results not only align with previous findings but also provide a computational perspective for explanation. Finally, shifting our focus to individual traits, numerous studies have shown that the level of social dominance significantly affects the decision-making process. Dominance is often defined within competition scenarios (Ligneul et al., 2017; Zink et al., 2008), and a higher level of dominance may correlate with better performance (Santamaría-García et al., 2014). Combining our findings with those of da Cruz et al. (2018), we offered a different perspective and concluded that social dominance can be considered a personality trait, emerging independently of social context. Moreover, through neuroimaging techniques, we identified neuromarkers associated with the dominance trait, potentially linked to the motivation driving response speed, irrespective of gender.

As a final note, the ability to learn and make decisions is crucial for navigating complex real-world scenarios (Collins & Shenhav, 2022), making it a fascinating area of study. I believe this understanding is key to unraveling the evolutionary process that has shaped us (Dukas, 1998).

8. References

The references listed here are those for the introduction (Ch.2) and discussion (Ch.7) chapters. The references for the studies are directly included in the corresponding chapter (i.e., sections 3.6, 4.6, 5.6 and 6.6)

Abohamza, E., Weickert, T., Ali, M., & Moustafa, A. A. (2020). Reward and punishment learning in schizophrenia and bipolar disorder. Behavioural Brain Research, 381, 112298. https://doi.org/10.1016/j.bbr.2019.112298

Anguera, J. A., Reuter-Lorenz, P. A., Willingham, D. T., & Seidler, R. D. (2011). Failure to Engage Spatial Working Memory Contributes to Age-related Declines in Visuomotor Learning. Journal of Cognitive Neuroscience, 23(1), 11–25. https://doi.org/10.1162/jocn.2010.21451

Aubret, A., Matignon, L., & Hassas, S. (2019). A survey on intrinsic motivation in reinforcement learning. arXiv Preprint arXiv:1908.06976.

Berlyne, D. E. (1950). Novelty and curiosity as determinants of exploratory behaviour. British Journal of Psychology, 41(1), 68.

Berridge, K. C., & Robinson, T. E. (1998). What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience? Brain Research Reviews, 28(3), 309–369. https://doi.org/10.1016/S0165-0173(98)00019-8

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. Cognition, 120(3), 322–330. https://doi.org/10.1016/j.cognition.2010.10.001

Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., & Efros, A. A. (2018). Large-scale study of curiosity-driven learning. arXiv Preprint arXiv:1808.04355.

Carver, C. S., Scheier, M. F., & Segerstrom, S. C. (2010). Optimism. Clinical Psychology Review, 30(7), 879–889. https://doi.org/10.1016/j.cpr.2010.01.006

Cheng, G. L. F., Tang, J. C. Y., Li, F. W. S., Lau, E. Y. Y., & Lee, T. M. C. (2012). Schizophrenia and risk-taking: Impaired reward but preserved punishment processing. Schizophrenia Research, 136(1–3), 122–127. https://doi.org/10.1016/j.schres.2012.01.002

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

Collins, A. G., Brown, J. K., Gold, J. M., Waltz, J. A., & Frank, M. J. (2014). Working memory contributions to reinforcement learning impairments in schizophrenia. Journal of Neuroscience, 34(41), 13747–13756.

Collins, A. G. E., & Shenhav, A. (2022). Advances in modeling learning and decision-making in neuroscience. Neuropsychopharmacology, 47(1), 104–118. https://doi.org/10.1038/s41386-021-01126-y

da Cruz, J., Rodrigues, J., Thoresen, J. C., Chicherov, V., Figueiredo, P., Herzog, M. H., & Sandi, C. (2018). Dominant men are faster in decision-making situations and exhibit a distinct neural signal for promptness. Cerebral Cortex, 28(10), 3740–3751. https://doi.org/10.1093/cercor/bhy195

Daniel, R., Radulescu, A., & Niv, Y. (2020). Intact Reinforcement Learning But Impaired Attentional Control During Multidimensional Probabilistic Learning in Older Adults. The Journal of Neuroscience, 40(5), 1084–1096. https://doi.org/10.1523/JNEUROSCI.0254-19.2019

Dubey, R., & Griffiths, T. L. (2020). Reconciling novelty and complexity through a rational analysis of curiosity. Psychological Review, 127(3), 455–476. https://doi.org/10.1037/rev0000175

Dukas, R. (1998). Cognitive ecology: The evolutionary ecology of information processing and decision making. University of Chicago Press.

Ghavamzadeh, M., Mannor, S., Pineau, J., & Tamar, A. (2016). Bayesian Reinforcement Learning: A Survey. https://doi.org/10.48550/ARXIV.1609.04436

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

Gruber, M. J., Valji, A., & Ranganath, C. (2019). Curiosity and learning: A neuroscientific perspective.

Hall, J. A., Coats, E. J., & LeBeau, L. S. (2005). Nonverbal behavior and the vertical dimension of social relations: A meta-analysis. Psychological Bulletin, 131(6), 898.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. Econometrica, 47(2), 263. https://doi.org/10.2307/1914185

Kidd, C., & Hayden, B. Y. (2015). The Psychology and Neuroscience of Curiosity. Neuron, 88(3), 449–460. https://doi.org/10.1016/j.neuron.2015.09.010

Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., & Fernández, G. (2009). Reinforcement learning signal predicts social conformity. Neuron, 61(1), 140–151.

Kuperberg, G., & Heckers, S. (2000). Schizophrenia and cognitive function. Current Opinion in Neurobiology, 10(2), 205–210.

Lighthall, N. R., Pearson, J. M., Huettel, S. A., & Cabeza, R. (2018). Feedback-Based Learning in Aging: Contributions and Trajectories of Change in Striatal and Hippocampal Systems. The Journal of Neuroscience, 38(39), 8453–8462. https://doi.org/10.1523/JNEUROSCI.0769-18.2018

Ligneul, R., Girard, R., & Dreher, J.-C. (2017). Social brains and divides: The interplay between social dominance orientation and the neural sensitivity to hierarchical ranks. Scientific Reports, 7(1), 45920. https://doi.org/10.1038/srep45920

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. Nature, 412(6843), 150–157.

Pietschmann, M., Endrass, T., Czerwon, B., & Kathmann, N. (2011). Aging, probabilistic learning and performance monitoring. Biological Psychology, 86(1), 74–82. https://doi.org/10.1016/j.biopsycho.2010.10.009

Rowell, T. E. (1974). The concept of social dominance. Behavioral Biology, 11(2), 131–154.

Salthouse, T. A. (2009). Decomposing age correlations on neuropsychological and cognitive variables. Journal of the International Neuropsychological Society, 15(5), 650–661. https://doi.org/10.1017/S1355617709990385

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. Science, 275(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

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

Sharot, T. (2011). The optimism bias. Current Biology, 21(23), R941–R945. https://doi.org/10.1016/j.cub.2011.10.030

Sidanius, J., & Pratto, F. (1999). Social Dominance: An Intergroup Theory of Social Hierarchy and Oppression (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139175043

Strauss, G. P., Frank, M. J., Waltz, J. A., Kasanova, Z., Herbener, E. S., & Gold, J. M. (2011). Deficits in positive reinforcement learning and uncertainty-driven exploration are associated with distinct aspects of negative symptoms in schizophrenia. Biological Psychiatry, 69(5), 424–431.

Strauss, G. P., Waltz, J. A., & Gold, J. M. (2014). A Review of Reward Processing and Motivational Impairment in Schizophrenia. Schizophrenia Bulletin, 40(Suppl 2), S107–S116. https://doi.org/10.1093/schbul/sbt197

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (Second edition). The MIT Press.

Suzuki, S., Harasawa, N., Ueno, K., Gardner, J. L., Ichinohe, N., Haruno, M., Cheng, K., & Nakahara, H. (2012). Learning to Simulate Others' Decisions. Neuron, 74(6), 1125–1137. https://doi.org/10.1016/j.neuron.2012.04.030

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and Uncertainty, 5(4), 297–323. https://doi.org/10.1007/BF00122574

van de Vijver, I., & Ligneul, R. (2020). Relevance of working memory for reinforcement learning in older adults varies with timescale of learning. Aging, Neuropsychology, and Cognition, 27(5), 654–676. https://doi.org/10.1080/13825585.2019.1664389

Waltz, J. A., Frank, M. J., Robinson, B. M., & Gold, J. M. (2007a). Selective reinforcement learning deficits in schizophrenia support predictions from computational models of striatal-cortical dysfunction. Biological Psychiatry, 62(7), 756–764.

Waltz, J. A., Frank, M. J., Robinson, B. M., & Gold, J. M. (2007b). Selective reinforcement learning deficits in schizophrenia support predictions from computational models of striatal-cortical dysfunction. Biological Psychiatry, 62(7), 756–764.

Watkins, C. J., & Dayan, P. (1992). Q-learning. Machine Learning, 8, 279–292.

Xu, H. A., Modirshanechi, A., Lehmann, M. P., Gerstner, W., & Herzog, M. H. (2021). Novelty is not surprise: Human exploratory and adaptive behavior in sequential decision-making. PLOS Computational Biology, 17(6), e1009070. https://doi.org/10.1371/journal.pcbi.1009070

Zink, C. F., Tong, Y., Chen, Q., Bassett, D. S., Stein, J. L., & Meyer-Lindenberg, A. (2008). Know Your Place: Neural Processing of Social Hierarchy in Humans. Neuron, 58(2), 273–283. https://doi.org/10.1016/j.neuron.2008.01.025

9. Curriculum Vitae

Wei-Hsiang LIN

Address : Chemin des Lentillières 3A, 1023, Crissier, Switzerland
Email: weihsianglin0214@gmail.com
LinkedIn: https://www.linkedin.com/in/wei-hsiang-lin/
Phone : +41783355615

Strengths

- Advanced knowledge of **machine learning** and **statistical** techniques
- Proficient in **data analysis** and **visualization** using **Python** and **MATLAB**
- Expertise in analyzing **large-scale neuroimaging** data such as **EEG** or **fMRI**
- **Project management** proficiency
- Expertise in **reinforcement learning**

Education

Brain Mind Institute,
Swiss Federal Institute of Technology, Lausanne (EPFL) — Lausanne, Switzerland
Ph.D. candidate (Doctoral assistant) — 08/2019 – 03/2024

Institute of Neuroscience, National Yang-Ming University (YMU) — Taipei, Taiwan
M.S. in Neuroscience — 09/2012 – 07/2014

Department of Information management, National Central University — Taoyuan, Taiwan
B.S. in Information Management — 09/2008 – 06/2012

Work Experience

Institute of Neuroscience (YMU) — Taipei, Taiwan
Research Assistant — 06/2018 – 05/2019
Designed, implemented, and troubleshot human fMRI experiments to investigate the neural mechanisms underlying decision-making. Skillfully modeled and analyzed behavioral and fMRI data using Generalized Linear Model (GLM) and Support Vector Machine (SVM), achieving high precision and deriving insightful results. Achieved recognition by having the results published in a high-impact scientific journal (PLoS Biology, IF = 9.5) and in a leading media report in Taiwan.

Institute of Neuroscience (YMU) — Taipei, Taiwan
Research Assistant — 12/2015 – 06/2018
Acquired wet lab experience by utilizing in-vivo two-photon microscope to explore spatial memory in mice. Analyzed behavioral and neuroimaging data, employing advanced image processing techniques in Matlab. Innovatively designed essential experimental compartments in collaboration with local enterprises, saving half the budget compared to importing from abroad.

Military services - *Armourer corporal*
Directed and supervised a team of five in weapons management and maintenance for a military camp of over 100 soldiers.

Taichung, Taiwan
11/2014 – 11/2015

Projects

Brain Mind Institute (EPFL)
Reinforcement learning in healthy aging and schizophrenia patients

Lausanne, Switzerland
08/2019 – 03/2024

- Designed experiments to investigate reinforcement learning abilities across varied populations (aging individuals, schizophrenia patients).
- Successfully applied the Q-learning model to data, capturing distinct behavioral properties across different subject populations.
- Engaged in bilateral cooperation with international teams (Georgia).
- Presented research findings at three international conferences and drafted three research papers. More details.

Behavioral and Neural Markers of Social Dominance

- Investigated how personality traits, especially social dominance, affect decision-making through experiments.
- Collected large-scale data (questionnaires and behavioral tests) from 150 participants through online platforms during COVID-19-related limitations.
- Employed correlation analysis and PCA for data analysis; conducted EEG recordings and analysis using ANOVA and mixed-effect models.
- Behaviorally, identified that individuals with high dominance respond faster in decision-making. Additionally, discovered a specific EEG pattern that could potentially serve as a neuromarker for social dominance.

Additional Experience

Teaching assistant

Lausanne, Switzerland
02/2021 – 07/2022

- Assisted in the development and implementation of a hybrid learning environment, moderating both online and offline teaching modalities.

Developing personal website

Lausanne, Switzerland
02/2022 - present

- Developed my personal website utilizing Hugo and JavaScript, showcasing my portfolio.
- Implemented tracking with Google Analytics to evaluate website traffic and user interaction data.

Cultural services

Taoyuan, Taiwan
06/2009 – 08/2009

- Organized and conducted a summer camp at two rural elementary schools, enhancing educational opportunities for local children.

Technical Skills

Computational Modelling: Reinforcement learning, decision theories, psychophysics

IT Proficiencies: Python, MATLAB, JavaScript, Tableau, Gorilla (online experiment platform), image processing techniques, EEG analysis using EEGLAB and Fieldtrip, fMRI analysis using FSL, MS Office Suite

Data Science Expertise: Statistical analysis, machine learning techniques, network analysis, cluster

analysis, data visualization

Project Management: Managed projects ranging from lab-based to clinical studies, handled diverse data structures from behavioral to neuroimaging, utilized tools like Obsidian and GitHub for efficient project management

Languages

English: Professional working proficiency | Chinese (Mandarin): Native speaker

Interests

- Serving as vice-captain of a table tennis team for two years. Participating in a baseball team as a player. Enjoying hiking and working out.
- Playing the violin for several years.
- Reading economic and financial magazines and engaging in investment activities.

Personal information

Status: Married | **Age:** 33 | **Residency:** Student permit B