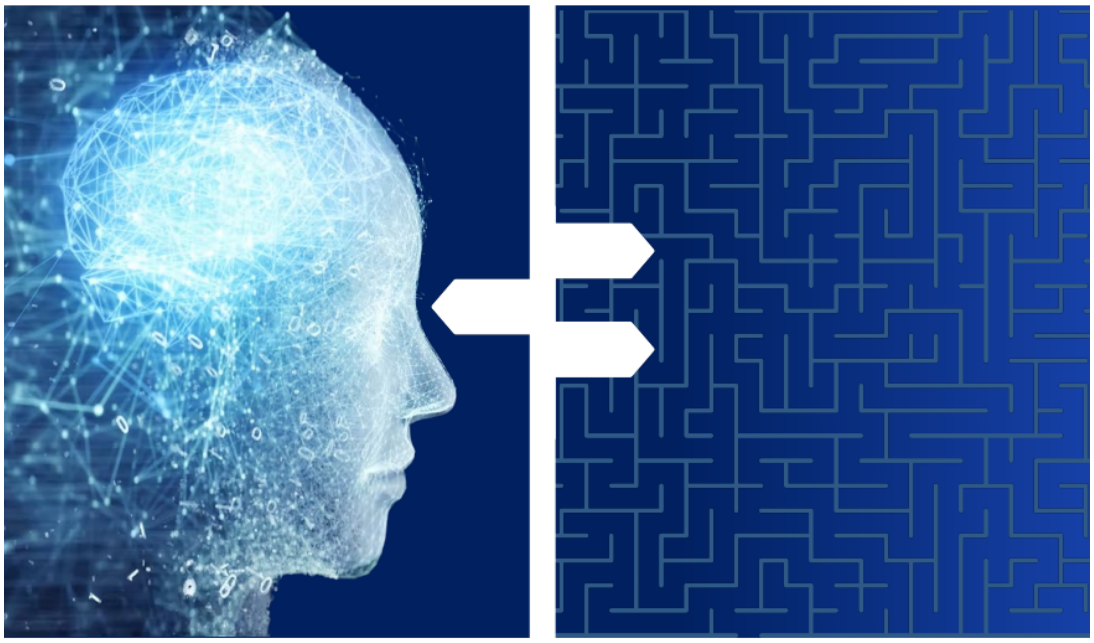
UNIVERSITAT DE BARCELONA

Decision Making in complex scenarios: a Reinforcement Learning Approach

Jeison Parra Tíjaro

DECISION MAKING IN COMPLEX SCENARIOS



A REINFORCEMENT LEARNING APPROACH

JEISON PARRA TIJARO

PHD THESIS

Decision Making in complex scenarios: A Reinforcement Learning Approach

Jeison Parra Tijaro

Cognition and Brain Plasticity Unit

Department of Cognition, Development and Educational Psychology,

University of Barcelona

2022

Doctoral program in Brain, Cognition and Behaviour

Supervisor:

Josep Marco Pallarés

Acknowledgements

These are the thankful words of an individual.

This has been the journey of a dreamer, a family man, a man in love, a son, a traveller, and now a newlywed.

Throughout this trip, I have encountered numerous hurdles, responsibilities, and duties, which have caused my academic work to falter on various occasions. A doctorate requires significantly more than experimenting, writing, reading, evidence, ongoing review, and results. This scholarly endeavour tugs at your every fibber, transporting you from the world of abstraction and allowing you to navigate an ethereal vessel to bountiful areas. Therefore, the most important aspects of pursuing a doctorate are internal to the student; how do you balance your life, your resources of time and money, your family time, and the time you devote to the people you love, with constant doubts about when you will be able to reach the other side of the coast.

I have never been alone, and I have always had the support and encouragement of great people in every situation and at every step. Before anything else, I'd like to thank my wife, who has become the most important person in my life, my drive and my source of energy to carry out this project. Thank you, Dimitra Vappa, for trusting in me and deciding on a life at my side; even when things appeared to be complicated, you were there. This thesis is dedicated to you, the love of my life, because you have been my source of illumination and have filled me with love.

This is also the opportunity to acknowledge my parents, Juan Parra and Arnobia Tijaro, to whom I owe my life and my first steps as a professional and scholar. They have trusted me since I was a child and have taught me to place importance on what is important. They influenced who I am now by their example. To Juan Parra, my dear old guy and friend, thank you for teaching me to be a warrior and to never give up and to always put my loved ones first, even if it means sacrificing something for the greater good. Our loved ones always have and will always be the north. To Arnobia Tijaro, thank you for your admiration, your devotion, and for always striving to be better for us; you have been our family's bulwark and its strength. I admire my parents, who accompany me everywhere I go. I would also like to thank my sister, Francy Helena, my brother-in-law, Diomedes, and my nephews, Andres and Samuel, who have always had me in their hearts and been there for me no matter the distance. They have allowed me to play a role in their decisions and be a part of their daily lives, despite the time difference and physical

distance. One of my primary sources of inspiration has been serving as a model for you.

To my Greek family, who welcomed me, gave me a hug when I needed it most, and opened the doors of their house and their hearts to me like another son. Mrs Eleni and Mr Meletis, thank you for everything and for being a part of my life. To my in-law sisters Evi and Konstantina, thank you for so many moments of laughter and understanding; you have always brought enchantment into my life, and I am privileged to be in your hearts.

Without the unwavering support of my dear Josep Marco Pallarés, a person from whose professional and personal example I have learnt, this would not have been possible, both academically and in my adaptation to the institution and city. You are the best supervisor I could have ever imagined; I will be eternally grateful to you for your patience and leadership in carrying out this project.

When I arrived in Barcelona with a bag full of aspirations, David and Johan embraced me as one of their own and assisted me in establishing my position within the institution and the department. I'd also like to thank Italo and Alberto, who were my teammates at the university for countless days and moments. Also I thank Unai who gave me a hand in the last stretch.

In addition, I would like to acknowledge Professor Ulrik Beierholm of the University of Durham, who hosted me for six months, was a fantastic facilitator for the implementation of my computational models and provided me with support before and throughout my doctorate stay. I thank you, and I hope we meet soon.

Lastly, I'd like to thank all the other PhD students with whom I've shared this trip and who have helped me along the way.

Sincerely thankful,

Jeison

Agredecimientos

Estas son las palabras de una persona agradecida.

Este ha sido el viaje de un soñador, de un hombre de familia, de un hombre enamorado, de un hijo, de un viajero, y ahora, de un hombre felizmente casado.

En este trayecto he enfrentado múltiples desafíos y he tomado un sinnúmero de retos, deberes y responsabilidades, que hicieron tambalear ni quehacer académico en más de una ocasión. Hacer un doctorado significa mucho más que la experimentación, la escritura, la lectura, la evidencia, la evaluación constante y los resultados. Esta empresa académica toca cada una de tus fibras, te lleva al mundo de lo abstracto y te permite conducir un navío etéreo hacia tierras fértiles. Así entonces, las cosas más importantes de hacer un doctorado transcurren en el interior del doctorando, cómo balanceas tu vida, tus recursos de tiempo y dinero, tu tiempo en familia y el tiempo que le dedicas a las personas que amas, con la duda constante de cuando podrás llegar al otro lado de la costa.

Para mi fortuna yo nunca he estado solo, y en cada momento y en cada escenario he contado con el apoyo y el soporte de personas extraordinarias. En primer lugar quiero agradecer a mi esposa, que se convirtió en la persona más importante, en mi motivación y mi fuente de energía para sacar adelante este proyecto. Gracias Dimitra Vappa por creer en mí y apostar por una vida a mi lado, incluso en esos momentos en que las cosas lucieron complejas, allí estuviste presente. Tú has sido mi luz y me has llenado de amor, y es por eso que esta tesis va dedicada a ti.

Este también es el momento te resaltar a mis padres Juan Parra y Arnobia Tijaro, a ellos les debo la vida y mis primeros pasos como profesional y académico, ellos confiaron en mí desde que era pequeño y me enseñaron a valorar las cosas importantes. Con su ejemplo me ayudaron a ser quien soy el día de hoy. A Juan Parra, mi querido viejo y amigo, gracias por enseñarme a ser un guerrero y nunca rendirme, priorizando siempre a las personas que amo incluso cuando hay que sacrificar algo por un bien mayor. El norte siempre ha sido y será el bienestar de nuestras personas amadas. A Arnobia Tijaro, gracias por tu adoración, por tu dedicación y por intentar siempre ser mejor para nosotros, tú has sido un bastión y nuestra fortaleza como familia. Los admiro mis padres y siempre van conmigo donde quiera que esté.

También quiero agradecer a mi hermana Francy Helena, mi cuñado Diomedes y mis sobrinos Andres y Samuel, que siempre me han tenido en su corazón y han estado siempre para mí sin importar la distancia, me han permitido tener un rol en sus vidas

y hacer parte de su día a día, sin importar el cambio horario o la lejanía física. Ser un ejemplo para ustedes ha sido una de mis principales fuentes de motivación.

A mi familia griega, quienes me acogieron, me dieron el abrazo cuando más lo necesitaba, me abrieron las puertas de su casa y su corazón como un hijo más, Mr. Meletis y Mrs Eleni, gracias por tanto y por estar presente en mi vida. A mis cuniadas Evi y Konstantina, Gracias por tantos momentos de diversión y comprensión, ustedes siempre han traído magia mi vida y me siento honrado de estar en su corazón.

En cuanto el apoyo tanto académico como en mi adaptación a la universidad y a la ciudad, esto hubiera sido imposible sin el incondicional apoyo de mi querido Josep Marco Pallarés, una persona de la que he aprendido desde su ejemplo como profesional y como persona. Te admiro profundamente y siempre estaré agradecido por tu paciencia y dirección para sacar este proyecto adelante, eres el mejor supervisor que jamás podría haber imaginado.

Para el momento en que llegue a Barcelona con una maleta cargada de sueños, David y Johan, Me acogieron como uno más de la manada y me apoyaron a establecer mi lugar dentro de la universidad y el departamento. También agradezco a Italo y Alberto que fueron mis coequiperos en tantos días y momentos en la universidad.

Hago también una mención especial para el profesor Ulrik Beierholm de la universidad de Durham, quién me acogió durante 6 meses y fue un gran facilitador para el desarrollo de mis modelos computacionales y me brindó su apoyo antes y durante mi estancia doctoral. te agradezco y espero nos encontremos pronto.

Finalmente agradezco a todos los demás doctorandos con los que compartí en el camino y que de alguna manera también hacen parte de este viaje.

eternamente agradecido,

Jeison

Table of Contents

Abstract.

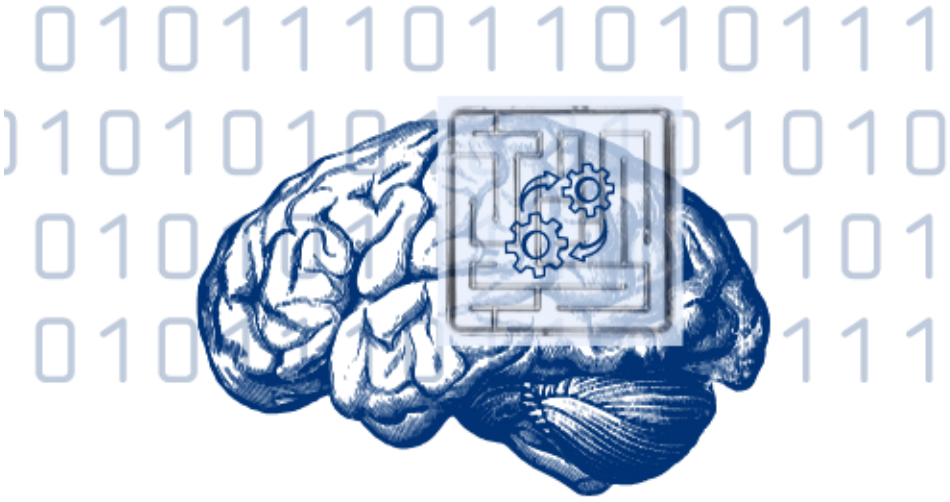
Our everyday actions are sequential and chained towards accomplishing goals. As an adaptive strategy, our tasks are divided and organised in stages leading up to an ultimate objective, which is specified by reaching progressive sub-goals. We create hierarchies in order to streamline our decision-making anytime an action has to be selected. Previous research has demonstrated that people chose alternatives with higher pseudo-rewards in order to achieve a sub-goal. As a behavioural strategy to reduce complexity, individuals break down routines in simpler stages and the completion of such intermediate states is reinforcing itself. Therefore, sub-goals are set hierarchically and their achievement act as pseudo-feedbacks that drives learning and influences decision-making. Still, this remarkable preference for pseudo-reward predictive stimuli has not been elucidated as a potential factor involved in sub-optimal choice behaviour. The goal of the present thesis is to **investigate the behavioural and neural correlates of pseudo-feedback processing in sub-optimal decisions**. To reach this goal, three studies are presented. In the first behavioural study, 226 university students participated in two experiments to test the hypothesis that pseudo rewards bias decisions. The task consisted of two alternatives, presented as two figures that were associated with different probabilities of obtaining pseudo-rewards. The results of this experiment revealed that people preferred the more pseudo-rewarding option even when this implied a reduction in the final global reward. In the second study, EEG was recorded from twenty-four healthy subjects who participated in a version of the two-step task used in the experiment 2 of Study 1. In particular, in this task, the probability of final reward decreased for the most-selected option (usually the one that provided more pseudo-rewards, as shown in study 1). Three different reinforcement learning approaches were used to model the behavioural data, and their results were used to study the oscillatory activity associated with reward and pseudo-reward prediction errors. Results showed that frontal theta oscillatory was associated with both reward and pseudo-reward prediction errors. Finally, in the third

study, we examined the role of the Ventral Striatum in reward- and pseudo-reward prediction errors during decision-making in sub-optimal settings. Nineteen university students participated in a functional Magnetic Resonance study performing a modified version of the two-step monetary task. Results showed that Ventral Striatum was involved in the computation of the prediction errors of feedbacks and pseudo-feedbacks and that the preference for the most pseudo-rewarding option was related to the activity of this area with pseudo-reward prediction errors. Overall, the three studies showed that the attainment of sub-goals is reinforcing and might bias decisions towards sub-optimal choices. In addition, we have demonstrated the critical involvement of the reward network (ventral striatum) and the theta oscillatory activity in hierarchical reinforcement learning.

Resumen.

Nuestras acciones cotidianas son secuenciales y están encadenadas para lograr objetivos. Como estrategia de adaptación, nuestras tareas se dividen y organizan en etapas que conducen a un objetivo final, el cual se consigue alcanzando metas secundarias progresivas. En este contexto, la creación de jerarquías comportamentales permite agilizar la toma de decisiones cada vez que se debe seleccionar una acción. Investigaciones anteriores han demostrado que las personas eligen alternativas con pseudo-recompensas más altas para lograr una meta secundaria. Como estrategia conductual para reducir la complejidad, los individuos descomponen las rutinas en pasos más simples y la finalización de tales estados intermedios es gratificante en sí mismo. Por lo tanto, los subobjetivos se establecen de forma jerárquica y su consecución actúa como pseudo-retroalimentación, lo cual impulsa el aprendizaje e influye en la toma de decisiones. Aún así, esta notable preferencia por los estímulos predictivos de pseudo-recompensa no se ha dilucidado como un factor potencial involucrado en el comportamiento de elección subóptimo.El objetivo de la presente tesis es investigar los correlatos conductuales y neuronales del procesamiento de pseudo-feedback en decisiones subóptimas. Para alcanzar este objetivo, se presentan tres estudios. En el primer estudio de comportamiento, 226 estudiantes universitarios participaron en dos experimentos para probar la hipótesis de que las pseudo recompensas sesgan las decisiones. La tarea constaba de dos alternativas, presentadas como dos formas que se asociaban con diferentes probabilidades de obtener pseudo-recompensas. Los resultados de este experimento revelaron que las personas preferían la opción con más pseudo-recompensa incluso cuando esto implicaba una reducción en la recompensa final global. En el segundo estudio, se registró EEG de veinticuatro sujetos sanos que participaron en una versión de la tarea de dos pasos utilizada en el experimento 2 del Estudio 1. En particular, en esta tarea, la probabilidad de recompensa final disminuyó para la opción más seleccionada (generalmente la que proporcionó más pseudo-recompensas, como se muestra en el estudio 1). Se utilizaron tres modelos de aprendizaje por refuerzo

diferentes para explicar los datos comportamentales y sus resultados se usaron para estudiar la actividad oscilatoria asociada con errores de predicción de recompensas y pseudo-recompensas. Los resultados mostraron que la oscilación theta frontal se asoció con errores de predicción de recompensa y pseudo-recompensa. Finalmente, en el tercer estudio, examinamos el papel del Estriado Ventral en los errores de predicción de recompensas y pseudo-recompensas durante la toma de decisiones en entornos subóptimos. Diecinueve estudiantes universitarios participaron en un estudio de Resonancia Magnética funcional realizando una versión modificada de la tarea monetaria de dos pasos. Los resultados mostraron que dicha estructura estuvo involucrado en el cálculo de los errores de predicción de feedbacks y pseudo-feedbacks, y que la preferencia por la opción que ofrecía más pseudo-recompensa estaba relacionada con la actividad de esta área con errores de predicción de pseudo-recompensa. En general, los tres estudios mostraron que el logro de los subobjetivos es un refuerzo y podría sesgar las decisiones hacia elecciones subóptimas. Además, hemos demostrado la participación crítica de la red de recompensas (en particular, el estriado ventral) y la actividad oscilatoria theta en el aprendizaje por refuerzo jerárquico.



Chapter 1. Introduction

Chapter 1 : Introduction

Decision-making is a vital skill for everyday living, adjustment to the environment, and autonomy, as it is the capacity to pick between two or more possibilities to accomplish one or more goals (Broche-Perez et al., 2016). Most cognitive activity can be related to a decision of some kind, and at a time, such association might be translated into actions. Therefore, brain functioning is related to behaviour, permanently striving to adapt to the environment and its changes (Morelli et al., 2021). Our behaviour, in general, depends on the decision-making processes that are carried out daily to cover our needs and goals. Considered from this point of view, the neuroscience of decision making (NDM) provides information on the brain mechanisms that underlie cognition during our choices, allowing a better understanding of the foundational elements of cognition and its impact on our behaviour (Doya, 2012; Marchau et al., 2019).

Most of the decisions that can be represented experimentally involve some observable behaviour that indicates that one alternative has been chosen over the others. In fact, most of the theory about the brain processes that underlie decision-making brain function comes from studying the neural substrate during the selection or planning of actions, either as perceptual responses to a stimulus or as a result of goal pursuance (Doya, 2012). Actually, decision-making is a broad topic that has been studied under multiple approaches, including philosophy, neuroeconomics, computational science, learning, and psychophysics. However, there is a giant distance between a single decision and a committed behavioural preference. Depending on the complexity of decisions, they might involve a variety of processes, including combining evidence over time, filling information threshold, giving value to different decisions, combining multiple sources of

information and learning associations, among many others. This might also yield to a variety of decision types. For example, value-based decisions emphasise the weighting of alternatives according to their expected utility (Polania et al., 2014; Machau et al., 2019), perceptual responses can explain cases that depend on the stimuli's features and the reaction to the environment, and phenomena such as commitment and complex learning are observed in the case of more elaborate behavioural routines (Diuk et al., 2013a).

Studies on valuation and reward processing have the potential to describe behavioural preference and learning strategies that maximise reward (Sutton and Barto, 2019). Therefore, the ability to weigh options based on the outcome is considered a sophisticated mechanism to approach desirable results. In addition, it is hypothesised that the brain mechanisms that support this reward processing also facilitate learning. In fact, many studies on decision-making have relied heavily on reinforcement learning (RL) theory. RL is based on calculations relating reward to behavioural 'state' and a number of alternative actions. For instance, while many decisions are reactive and almost automatic, individuals often spend time and energy thinking about the possible outcomes from different options and their relative value (Dulac-Arnold, 2019; Dayan, 2002). Under this vision, decision-making is refined due to learning through individual actions (Pateria et al., 2021; Deco et al., 2013).

However, this perspective continues to be limited to understanding complex actions, as it is the case of an activity that depends on a long sequence of steps (e.g., paying the bills online requires multiple steps from turning on the laptop to completing the payment). Yet, the brain representation of decision-making has been approached by involving more

primitive mechanisms and statistical learning. Therefore, decision-making requires the involvement of intricate brain networks underlying complex computational representations (Pateria et al., 2021; Rushworth et al., 2011; Botvinick et al., 2012).

Decision Making and Learning.

The concepts of learning and decision-making are inextricably intertwined. Typically, individuals are able to check into the precise result of their decision only after they have already taken it. These outcomes operate as insights and provide the individuals with a more detailed understanding of the features of their chosen alternatives. Therefore, this valuation helps to influence their subsequent decisions under the same or similar conditions (Gebhardt et al., 2021; Diuk et al., 2013a). Despite this inherent link between decision-making and learning processes, many of the decision-making theories do not contain or adequately stress a learning component (Pleskac, 2008; Hassall, 2019).

Individuals strive to find a way to optimise their chances of acquiring the resources necessary for survival while decreasing their chances of meeting conditions that may cause loss or damage. To address this, individuals develop a number of strategies, such as, as previously stated, accumulating evidence from previous interactions with the environment to build knowledge about their odds. This story of interchange allows the individuals to predict and examine new findings to enhance learning and upgrade their ability to deal with uncertainty. Therefore, different systems help monitor the results of decisions. These different systems coexist, as shown by various

neuroimaging studies (O'doherty et al., 2017; Decker et al., 2016). Among others, two main strategies to control behaviour have been identified: those in which actions are guided by the characteristics of the stimulus (i.e., cue learning) and those guided by the features of the reward (goal-directed behaviours; Decker et al., 2016). In the first case, the actions are determined by the reaction to the occurrence of an external stimulus (stimuli driven), while the goal-directed behaviours seem to be motivated towards a specific reward (Nachum et al., 2018; Valentin et al., 2007). These strategies have also been distinguished and are dissociable, referred to as habitual decision-making and objective-oriented decision-making. These two distinct sorts of evaluative processes can influence one's choices and are distinguished and represented differently by theoretical models (Daw et al., 2005; 2011). Stimuli-driven behaviours have been sometimes referred to as habitual behaviours where monitoring and error detection systems are not so embedded as in goal-directed behaviours (Decker et al., 2016).

In any case, the stimuli-driven strategy has been linked to decision-making and statistical learning (Polania et al., 2014; Mormann and Russo, 2021). In this theory, the individual responses highly depend on the magnitude and the salience of a stimulus, and consequently, this response might be matched to a reward. In a stimulus-driven behaviour, decision-making occurs more quickly and automatically and would be closer to primitive adaptive mechanisms. Hence, under this strategy, the interchange with the environment is more reactive, making it very difficult to be used in more complex behaviours where the salience of the stimuli is not the central factor.

On the other hand, a goal-guided strategy proposes a proactive approach to foraging the exchange between individuals and their environment. Therefore, regarding this strategy, individuals seek the best options and tend

to repeat those actions associated with rewards and reduce those actions that move them away from a desirable result (Osband et al., 2019). Therefore, this learning strategy is progressive and allows the control of responses, associating decision-making with the value given to the reward and not to the stimulus. A slower goal-directed approach examines possible actions and their expected outcomes, while a faster and more automatic habitual process connects rewards with cues, allowing for the reflexive repeating of previously successful activities (O'Doherty et al., 2017). More interestingly, these strategies (stimulus-driven and goal-directed) appear to work together and coexist, as shown in rodents (Dickinson et al., 1995) and human (Valentin et al., 2007) experiments. In fact, their parallel presence in learning and decision-making may indicate that both complement each other to facilitate access to resources and meet survival needs. Therefore, in both strategies, individuals use adaptive mechanisms to being able to reduce uncertainty and make beneficial decisions.

As explained, individuals pursue goals to increase their benefits, and this ability to direct behaviour towards goals requires the capacity to track rewards, their likelihood of occurrence, and additional related characteristics (e.g., risk, magnitude, valence). The term "value" is used in the literature to refer to the manner in which these characteristics are integrated. Therefore a *'value function'* is a way of operationalisation of such adaptive strategy, where we can foresee a potential connection with RL schemes and computational modelling such as temporal difference. In the literature, different types of value functions have been proposed. Among the most used, some of them are linked to the value of the available actions, $Q(s, a)$ (being s the available states and the actions, see below) and others associated with the value of achieving certain state V(s), also known as state value function. Both the action selection and the value function are determined by the total

amount of rewards after achieving a final state or receiving feedback (Sutton and Barto, 1998; 2018). This idea resembles the choice theory (Gardner et al., 2019; Shiv et al., 2005), which proposes that the anticipated value of an option or action is an additive function of its various reward outcomes, which means that it represents learning through a record of previous choices. Along with the choice behaviours, every option is weighted by its chance of occurrence. This approach to choice-based behaviours is coherent with the "prospect theory" developed by Kahneman and Tversky, which was successful in predicting a variety of human behaviours. Particularly, the prospective theory explains the deviance from the predictions of classical decision theory, and it provided a strong impetus for research on value-based decision making (O'Doherty, 2017; D'Acremont and Bossaerts, 2008; Glöckner and Pachur, 2012; Takahashi, 2012). This connection is evidence that Learning and Decision Making are intertwined in a permanent bias yet not always clarified. However, neuroscience and computational models have shed light on the neural and behavioural substrates that compound such adaptive mechanisms encompassing decision-making and learning.

Reinforcement Learning (RL)

When we think about the nature of learning, the notion that we learn through interacting with our environment is usually the first that comes to mind. Even though it can be approached as an abstract concept, there have been diverse computational efforts to address it from a more measurable perspective. RL constitutes one of the most widespread and well-founded approaches to addressing instrumental learning, especially in explaining the

interactions of an agent with the environment. RL describes learning in an incremental way, where individuals collect evidence on a trial-and-error basis to get closer to a desirable reward (Dulac-Arnold et al., 2019). RL aims to understand goal-directed learning and decision-making in a specific time frame through computational methods. Unlike other formal models, RL seeks to describe the interaction of an agent with its environment in a deterministic way, giving a predominant role to the agent's choices and attempting to maximise future rewards. Therefore, the agent is not just a passive learner but a decision-maker. In fact, the conceptualisation of a more active learner constitutes a pioneering method in addressing computational difficulties such as learning involving long-term goals acquisition (Sutton and Barto, 1998). Moreover, RL constitutes a theoretical, scientific, and computational framework that defines the interaction of an agent with the ability to learn with its environment in terms of states, actions and rewards (Sutton and Barto, 1998; Dayan, 2002; Moerland et al.; 2020).

Indeed, RL focuses on the agent's actions concerning its environment, considering that the environmental variables could change over time along with its own expectations of the reward. Then, goals and actions are valued every time the agent exchanges with the environment and weighted in relation to previous decisions. Therefore, this approach is based on goal-directed behaviour, where the agent consciously strives for a desirable outcome. As a formal method, RL attempts to explain learning by considering *the characteristics of the environment (states), the behaviour motivated by an expectation (actions) and the impact of the behaviour chosen to obtain a result (rewards,* see Figure 1, Sutton and Barto, 1998*).* These characteristics imply an awareness of causality between actions and rewards, an intention to reduce uncertainty, and the presence of declared goals that guide behaviour and expectation. RL uses a rational approach,

beginning with a fully developed, interactive, goal-seeking agent trying to maximise its gain (Solway and Botvinick, 2012; Levine, 2018).

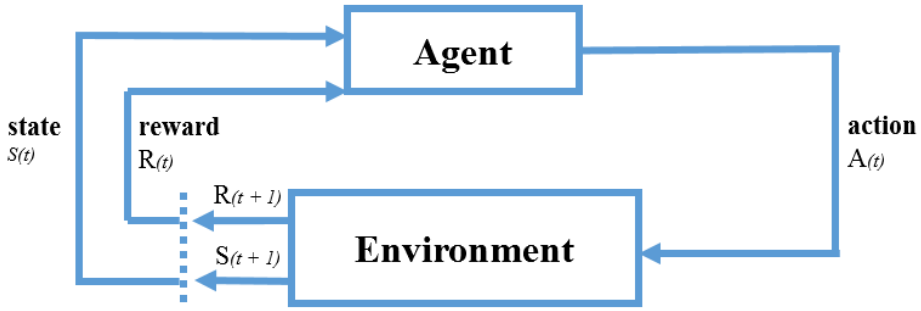


Figure 1.1 Representation of a standard RL model, where the agent has full access to the reward history and learns continuously and cyclically. This scheme is known as the Markovian Decision Process (MDP). These models propose a rational agent in an environment in which outcomes might be probabilistic but also depend partially on its decisions, which improves its capacity to access rewards permanently and refines its actions. The agent increases its ability to take benefits from the environment in a progressive manner by collecting evidence from previous actions (Botvinick, 2009). Reproduced from Sutton and Barto (2015) with permission.

Therefore, in the RL framework, the agent is the component that determines what action to perform. The agent can use any observation from the environment and any internal rule that it possesses to pick an available action. To clarify, the agent could be a human individual, animal, robot, a local network or any decision-maker capable of learning (Sutton and Barto, 2018; Osband et al., 2019). The mapping between states and actions is known as policy. After the first state is provided by the environment to the agent, internal rules are updated following a Markovian chain. As a result, the agent processes each state using a policy function that determines what action to take based on the value of each action. Yet, this policy function is updated every time the action consequence is resolved. In addition, RL methods are determined by how an agent responds to a reward signal and

feedback-related cues (information available in the environment). Therefore, it is essential to understand how to optimise goal-directed behaviours and maximise the desirable reward in the future (refinement process). To accomplish this, the agent gathers information about the actions taken in the past and uses it to improve the policy for the future. Furthermore, RL models are based on the conception of scalar learning, where primary information comes from the difference between the expected rewards and the actual repercussions of the agent's actions. This contrast, known as Prediction Error (PE), is the basis of the behavioural refinement process and leads to the attempt to close that gap between what is expected and what is obtained in a new state (Botvinick, 2012). PEs are commonly considered the engine of learning, as they are used to update expectations in order to make predictions more accurate. Thus, the speed of adjustment to the PE and the improvement in the accuracy of the decisions might be a critical indicator to describe the agent's learning process. When prediction errors are refined and adjusted in relation to a reward and its representation and expectation, they are known as Reward Prediction Errors (RPE).

As mentioned, RL models include the interaction between a decision-maker and its environment, in which the agent attempts to achieve a goal despite environmental uncertainty. Therefore, RL is also a model capable of describing a process of adaptation to environmental demands, gaining certainty about the effectiveness of their actions to access favourable results (Dayan and Balleine, 2002; Moerland et al.; 2020). In fact, the agent's actions might include altering the future state of the environment, thus influencing the available options and possibilities in the future. For instance, a driver that explores a new route to get to the same destination or a surgeon that explores a new procedure to retire a subdermal tumour are examples of agents that influence the action selection process. Thus, decisions gain

precision as a result of estimating the consequences of actions, which may involve foresight or preparation. At the same time, RL assumes the results of the actions cannot be fully predicted in any case; consequently, the agent must frequently check its environment and react appropriately in order to adjust its behaviour.

The Markovian Decision Process (MDP)

As expressed before, MDP is a representation of the environment's "dynamics", or how the environment will react to probable actions taken by the agent in a particular state. More formally, an MDP is supplied with a *transition function*, which outputs a probability of moving to any of the subsequent stages given the present state and an action taken by the agent among the pool of available actions (Otterlo and Wiering, 2012). Additionally, an MDP has a reward function. The reward function, intuitively, generates a reward or punishment in response to the present state (and, possibly, an action taken by the agent and the next state of the environment). The combination of such functions is referred to as the *model of the environment*, and the whole knowledge of this model might allow the agent to foresee an accurate representation of the upcoming outcomes. So, the MDP is the problem that an agent needs to address through learning strategies and the policy that predicts future rewards is the solution (Dayan and Niv, 2008).

Nevertheless, in most cases, particularly the naturalistic ones, the agent does not have accumulative evidence to have a full model of the environment (MDP associated with the environment). As a result, the agents cannot

accurately estimate a policy, which is uncertain and depends on the further exploration and exploitation of the available actions and the immediately previous and future outcomes. Notably, the main difference remains in the access to the transition and reward functions associated with the environment. The sooner these functions are estimated and approximated, the sooner the agents can use them to find an optimal policy (Sutton and Barto, 1998; Osband et al., 2019). To determine the best strategy, the agent must observe them and interact with the environment. This *reinforcement learning problem* requires that the agent must estimate a policy by refining/updating its assumptions about the environment's dynamics. The agent gradually gains an understanding of how the environment reacts to its actions and may thus begin to estimate an optimal policy.

Model-based vs Model-free algorithms.

Two basic categories of RL models have been proposed, the Model-free RL and Model-based RL (Daw et al., 2005; Daw, 2015; Decker et al., 2016). In model-free RL, trial and error outcomes are used to adjust an action value related to a stimulus, taking observations directly from experience without estimating a whole map of the environment (model of the world/environment). This method encourages the re-enactment of previously learned behaviours, this is, repeating every action that derived in a positive reward without or just partially knowing the transition function and reward function. Therefore, the agent follows his trial-and-error experience for setting up the optimal policy, which might combine possibly erroneous estimates or beliefs about state values since the model of the environment is not fully known (Dayan and Niv, 2008). Model-based RL, on the other hand, uses a flexible but computationally intensive procedure to make

judgments by assigning a "decision tree", which consists of links between state transitions and outcomes (known as the transition function and the reward function, Huang et al., 2020). In Model-based RL, the agent builds an internal model of the environment's transitions and the imminent outcomes based on available information on the possible actions. Therefore, model-based RL is more statistically efficient than model-free RL, but both accomplish the general purpose of any RL model, which is policy optimisation.

In this context, a model-based RL estimates the best policy by utilising a model of the environment. The agent has access to an approximation of the transition and reward functions, which have been learnt from the environment or provided to the agent. In general, because the agent knows an estimate of the transition function, it may theoretically forecast the dynamics of the environment faster and more accurately than in model-free RL. However, it is worth noting that the agent's transition and reward functions enhance his estimate of the optimum policy by approximations of the "real" functions. In contrast, a model-free RL determines the best policy relying on guessing the environment's dynamics (transition and reward functions). In practice, a model-free RL algorithm either estimates a "value function" or a "policy" directly from experience (the agent's interaction with his environment) without employing a prior knowledge or learning phase that provides a model of the environment.

To summarise, the conceptualisation of model-free and model-based RL depends on whether the agent employs predictions about the environment's reaction when learning or behaving. The agent can use a single prediction from the model of the future reward and state (model-free), or it can query the whole model for the predicted next reward or the whole distribution of

next states and next rewards (model-based). These predictions can be generated completely independently of the learning agent, for example, by a computer code or another agent that knows the rules of the environment. For instance, in a board game, an agent would go through model-free RL when evaluating a policy from his own interaction with the game; this is if his plays succeeded or not. Instead, in model-based, the agent would have a full or approximate understanding of the rules behind the game. Some standard RL approaches propose a division between two components of a learning system or agent: an *actor* that selects actions based on their weighted associations with the current state of the world (policy) and a *critic* that generates an estimate of the reward associated with the environment (value function). Both the actor's policy and the critic's value function are learnt via experience. However, standard versions of these actor-critic schemes seem to fail in predicting learning in more complex situations. We will take a closer look at the actor-critic approach when discussing the hierarchical reinforcement learning, elucidating how this combination can outline the existence of different systems during action selection and feedback processing.

Reinforcement Learning (RL) and its link to Decision Making (DM).

RL is an approach to describe the adaptation in which an agent uses prior experience to enhance future results. The RL theories have contributed to the development of value-based choice theories, therefore, have shed light on the bridge between Learning and DM, given its capacity to operationalise the relationship between individuals and the environment. As stated above, in

RL, actions are chosen based on their *value functions*, which represents the predicted future reward associated with each available action. The exposition of punishments and rewards and the role of the environment in every interaction may affect these values, updating them to be used in the future. Thus, decision-makers refine their choice by collecting evidence and using it to adjust the policy, the value of each option and the expected value of rewards (Lak et al., 2020; Lee et al., 2012; Niv, 2009). How individuals value different options and pick their choice among a pool of options has been in the spot of different disciplines such as economics, psychology, and neuroscience. Tracing back to the Decision-Making concept, among others, there have been two main traditions that have hoarded the attention; these are the *normative perspective* (von Neumann and Morgenstern, 2007) and the *prospect theory* (Kanheman and Tversky, 1983; Solaki et al., 2021). Economical normative theories emphasise the concept of utility, where individuals always choose the action with the best return in benefits. This approach considers that after identifying well-defined criteria, individuals can determine the most optimal course of action (Malecka, 2020; Baron et al., 2004).

Nevertheless, this framework fails to explain more naturalistic schemes where the environment and the expectations change constantly. Furthermore, an idealised set of conditions with all the information revealed, prospect theories bring the idea of a decision-making process based on heuristics, which points out an individual who self discovers how the environment operates by trial and error. This approach is particularly efficient in describing an agent gaining control over uncertain conditions while relevant information associated with rewards needs to be discovered (Kahneman and Tversky, 1979; Solaki et al., 2021). Besides these two perspectives being fundamental in the scientific exploration of the neural correlates of Decision

Making, none of them involves the issue of learning, which in theory must take place any time an agent develops a preference or establishes a choice behaviour.

For instance, normative-economic theories stand that individuals weigh their options regarding their utility, choosing any time the option with the greatest value, therefore, the most optimal one. Although this paradigm is especially successful when all the criteria involving the decision are known, it does not consider aspects such as exploration or sub-optimal decision making (Lee et al., 2012; van der Meer and Redish, 2010; Lak et al., 2020). It is clear that all individuals pursue to maximise their rewards, but from the normative perspective, the mechanisms underlying the value functions are either evolutionary aspects or determined by individual experience. In fact, individuals' mechanisms involved in adjusting for the immediate subsequent decision are overlooked and sealed as a non-observable variable (Lee et al., 2012). On the other hand, RL describes how an agent updates its value function anytime it enters in contact with the environment and gets an outcome from its actions. Therefore, RL provides greater scope to the NDM anytime the interaction between an agent and the environment influences choice behaviour.

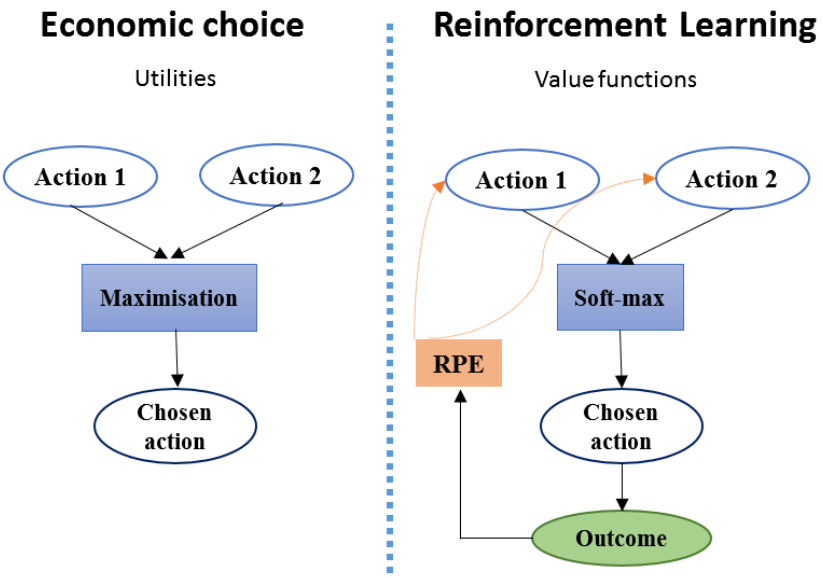


Figure 1.2. According to economic theories, decision-making refers to picking the action with the highest benefit. In reinforcement learning, actions are selected probabilistically (i.e., softmax) based on their value functions. In addition, value functions are modified based on the outcome (reward or punishment) of the agent's selected action. Reproduced from Lee et al. (2012) with permission.

As in the cases of the normative theories, RL's objective is to maximise future rewards. Similarly, individuals value the utility, but in RL, the value function lies on the expected value of the rewards and is usually estimated on the bases of trial and error. This standard RL idea is very efficient in understanding feedback processing when the reward is not so distant from the actions (Niv, 2009; FitzGerald et al., 2012; Adkins and Lee, 2021). This central idea of a value-based choice is not just pivotal for an RL agent exploring the world, but also it is coherent with the idea of an individual who reckons for optimality. Perhaps the most important insight of this convergence is that the value function happens in a specific time frame or temporal resolution where the delivery of the reward is expected. As shown

in the graph, the RL agent accounts for a heuristic to progressively discover the action with the greatest value as a result of collecting evidence.

Neural Correlates of Reward Processing.

Several studies have been conducted to ascertain the neurological substrates of value and reward-related behaviour (goal-oriented behaviour). While causal relationships are not always obvious due to the variety of components involved in such processes, decision-making can be separated into stages involving several brain regions. These stages are the creation of a representation, the choosing of options/action selection, the monitoring of such selections, and the processing of feedback (Morelli et al., 2021; Ernst, 2005). Importantly, the reward network plays a role in these stages. The reward circuit is composed of multiple cortical and subcortical areas that collaborate to mediate various elements of incentive learning, resulting in adaptive behaviours. This circuit is centred on the cortical–Basal Ganglia network. In this intricate network, the prefrontal cortex (PFC) and the Basal Ganglia are involved in learning and in the formation of goal-directed behaviours and action plans (Haber, 2017).

Therefore, while exerting a goal-directed strategy, the reward circuit is fundamental to building a behavioural reaction to external environmental stimuli, motivation and reward information related to a goal. In other words, merely desiring to succeed or access a reward is not enough to prompt learning. For instance, to play a tennis match, a player needs to understand the game's rules, the available options and so on. Thus, developing action plans aimed at achieving a goal requires a blend of reward processing,

cognitive, and motor control mechanisms. However, cortical–Basal Ganglia processing models have stressed the separation of functions, emphasising distinct and parallel pathways such as error detection and statistical encoding of available information.

For instance, Schultz demonstrated (1998) that dopaminergic neurons of the Ventral Tegmental Area (VTA) compute not just reward but also reward prediction error. VTA is located close to the substantia nigra in the midbrain. Although it contains a variety of different types of neurons, it is primarily defined by its dopaminergic neurons, which project from the VTA throughout the brain, and are particularly well-known for their primary link to the Nacc/VS and amygdala, serving as a critical component of the reward neural network.

Importantly, Schultz et al. (1998) demonstrated that when a stimulus is reliably connected with a reward (e.g., in a classical conditioning setting), VTA neurons initially fire in response to the reward delivery, but later they fire in response to the predictive or conditioned stimulus. Alternatively stated, VTA neurons encode the history and the prediction of the reward. Additionally, when the experienced outcome deviates from what is expected, the same neurons activate. This signal is referred to as prediction error (which is central in the computation of reinforcement learning models), and it is capable of encoding both positive and negative deviations from the anticipated outcomes (positive and negative prediction errors). This evidence paved the road for a collaborative research commitment spanning the domains of neurophysiology, neurobiology, and psychology to elucidate the role of VTA input to cortical and subcortical brain regions in the neural representation of value and reward prediction (Haber and Knutson, 2010; Haber, 2017; Silvetti, Fias, and Verguts, 2014).

While reward-responsive cells are found throughout the brain, the cortical–basal ganglia circuit lies at the heart of the reward system. The medial prefrontal cortex (mPFC), the ventral striatum (VS), the ventral pallidum (VP), and the midbrain dopamine neurons have been proposed to be critical regions in this network. Additionally, other regions, such as the dorsal prefrontal cortex (dPFC), lateral habenular nucleus, amygdala, and thalamus, contribute significantly to the regulation of the reward circuit (Harber, 2017). Through connectivity between these locations, a complex neural network is formed that is topographically ordered, ensuring functional continuity along the corticobasal ganglia circuit. However, the reward circuit is not self-contained. Additionally, the network has particular locations where convergent pathways offer an anatomical basis for functional domain integration (Haber, 2017).

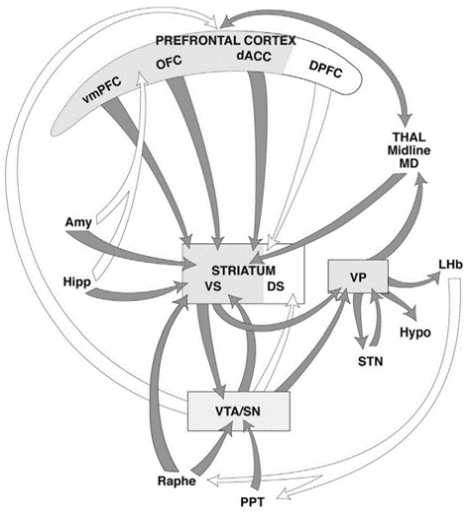


Figure 1.3. Connections of the main components of the reward circuit. The grey-highlighted areas reflect the regions' basic linkages. Amygdala; dACC, dorsal anterior cingulate cortex; DPFC, dorsal prefrontal cortex; DS, dorsal striatum; hypo, hypothalamus; LHb, lateral habenula; MD: mediodorsal thalamic nucleus; OFC, orbital frontal cortex; PPT, pedunculopontine nucleus; STN, subthalamic nucleus; VTA, ventral tegmental area; SN, substantia nigra. Reproduced from Haber (2017) with permission.

A key structure of this network is VS. The VS is critical for appetitive behaviour and reward anticipation (Huang et al., 2020), encoding value at the time of reward processing and promoting value-based learning (Diekhof et al., 2012). Striatal activity is often correlated with individual preferences and the value placed on a particular stimulus by participants (Castegnetti et al., 2021; Levy and Glimcher, 2011; Sabatinelli et al., 2007).

Additionally, the mPFC has been linked to several elements of value processing, such as reward prediction and outcome value coding. Indeed, in a meta-analysis by Oldham et al. (2017) with 22 neuroimaging studies using monetary incentive tasks, they studied the brain substrates of reward and loss processing, specifically to determine whether these processes were governed by a distinct or generalised system or a distinguishable one. The authors found that both reward and loss anticipation stimulated the striatum, thalamus, amygdala, and insula, showing that a generalised system is activated during this processing phase (Figure 1.4). This mechanism almost certainly plays a critical part in creating motivated responses that allow for the achievement of an optimal outcome.

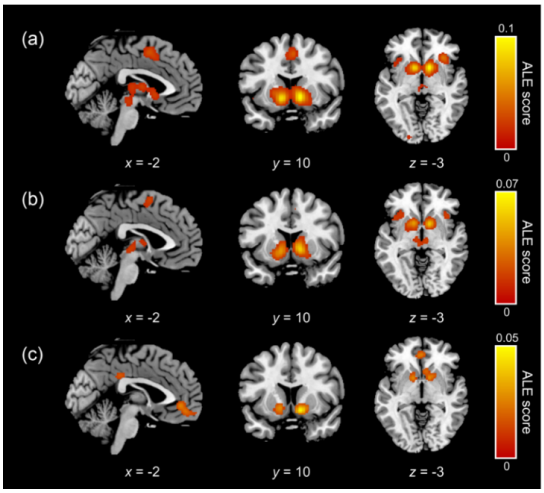


Figure 1.4. Meta-analysis on (a) reward anticipation, (b) loss anticipation and (c) reward outcome (consummatory phase). Reproduced from Oldham et al. (2017) with permission.

Unlike some earlier studies, the findings did not implicate the OFC/vmPFC during anticipation, suggesting that it only participates in reward anticipation when several choices are available (not binary paradigms and/or go-no go tasks) or when outcomes are extremely certain, in this case, a more model-free RL algorithm for instance. These results from the ALE meta-analyses aligned with previous findings where passive rewards and reward outcomes were linked to the role of the ventral striatum during reward processing (Diekhof et al., 2012; Daw, 2015). Functional neuroimaging studies have also demonstrated magnitude and likelihood signals associated with expected outcomes in the human striatum (Haber and Knutson et al., 2010) and midbrain (D'Ardenne and Hennigan, 2015), paralleling the results from single-cell dopaminergic neurons (Schultz, 1998), which increase activity in proportion to both the amount and likelihood of the expected reward. While the aforementioned fMRI studies support the assumption that these regions encode a single choice variable that encompasses both probability and magnitude (Palminteri et al., 2015),

some studies have revealed a spatial breakdown of probability and magnitude information within the striatum (Kahnt and Tobler, 2017).

Different brain areas have been implicated in the coding of value in model-based (goal-directed) and model-free behaviour. Within the striatum, separation of these distinct value signals has been demonstrated, with activity in the caudate nucleus coding for model-based value signals (Chiu et al., 2017) and activity in the putamen coding for model-free value signals (Lee et al., 2014). Both areas, however, are functionally connected to the mPFC, where they can be combined to facilitate value-based decision-making. This is consistent with the aforementioned findings demonstrating activity associated with value during value-based choosing (Kahnt and Tobler, 2017). Importantly, depending on the task settings and the characteristics of the stimuli, different neural correlates could be elicited (Wunderlicha et al., 2009; Lee et al., 2012). Therefore, Multiple brain areas, including the posterior parietal cortex, have been connected with neural activity related to action-value functions, according to research. (Platt and Glimcher 1999; Gold and Shadlen, 2007; Padoa-Schioppa et al., 2017; Noppeney, 2021), dorsolateral prefrontal cortex (Morris et al., 2014, Lin et al., 2020) and premotor cortex among others (Lüscher and Janak, 2021; Gremel and Costa, 2013; Cai et al. 2011). Additionally, the state value function has also been identified and tracked in various brain regions, being related to the activation of the posterior cortex and the dorsal striatum (Seo et al., 2009; Cai et al., 2011; Schultz, 2013). Distinctly, neurons that encode state value functions are also located in the ventral striatum (Kim et al., 2009; Cai et al., 2011), anterior cingulate cortex (Seo and Lee, 2007; Kolling et al., 2016; Soltani and Izquierdo, 2019), and amygdala (Seo and Lee, 2007; Jenison et al., 2011; Malvaez et al., 2019).

Oscillatory Activity.

As stated in previous sections, the reward network is crucial for learning and decision-making processes. The VTA, the ventral and dorsal striatum, and the ventromedial prefrontal cortex, among others, have been reported as components of reward processing mechanisms, therefore implied in learning from feedback encoding. To accomplish the goals of accessing limited resources, the different regions of the reward network interact with other areas, which are involved in cognitive and emotional processing (Watts and Bernat, 2019). In order to synchronise the different areas of the reward network and their association with other areas, some neural mechanisms have been proposed (Cavenagh et al., 2010). Among others, oscillatory activity seems to be a suitable mechanism for the communication of distant brain regions (Lega et al., 2011; Andreaou et al., 2017). Therefore, neural oscillations of variable frequency are a feasible candidate for the mechanism behind the reward system's flexible communication (Cohen, 2017).

Some electroencephalography (EEG) investigations have established that distinct reward-related stimuli elicit frequency-specific responses. First, researchers have identified that positive outcome processing is mostly connected with oscillations in the beta-gamma frequency during gambling tasks (Marco-Pallarés et al., 2015). Beta oscillatory activity is a time-frequency component (20–35 Hz), which is commonly recorded 200–600 ms after reward feedback at central and frontal midline locations (Glazer et al., 2018), although other studies find beta to be lateralised (Van de Vijver et al., 2011). Beta is sensitive to both performance and rewards evaluation during feedback processing, with research indicating that positive feedback increases beta power while negative feedback desynchronises its

signalling (Marco-Pallarés et al., 2009; Van de Vijver et al., 2011; Glazer et al., 2018). Gain-related beta activity is associated with activity in reward-related regions such as the ventral striatum (VS) and orbitofrontal cortex (Mas-Herrero et al., 2016). In addition, other studies have described beta power during reward anticipation, particularly during cue evaluation, implying a probable association with changes in dopaminergic activity (Apitz and Bunzeck, 2014).

On the other hand, theta oscillatory activity is linked to loss processing and reward prediction error (4–8 Hz; RPE; Glazer et al., 2018, Marco-Pallarés et al., 2015; 2016), with a typical, detected activity at frontal-midline regions. The majority of research studying midfrontal theta activity has focused on reward outcomes using typical measurement windows ranging from 200 to 500 ms after the exposition to a stimulus (Cohen and Cavanagh, 2011). Some evidence suggests that theta may also be involved in reward anticipation, particularly in connection to RPE encoding (Gruber et al., 2014; Cohen, 2017). Theta activity following feedback might serve as a mechanism for communication across distal brain areas within the same network, for instance, the reward pathways (Cavanagh et al., 2010). Midfrontal theta elicitation is sensitive to performance and reward evaluation, with increased power after receiving negative feedback (Bernat et al., 2015; Cavanagh et al., 2010), particularly regarding monetary losses, behavioural adjustments in response to losses (Cavanagh et al., 2010) and learning rate (Mas-Herrero and Marco-Pallarés, 2014).

Theta oscillations also play a role in the formation of Feedback Related Negativity (FRN; Marco-Pallarès et al., 2008). To support this view, converging evidence indicates that both the FRN and theta are generated in the medial prefrontal cortex, including the ACC (Gehring and Willoughby,

2002, Mas-Herrero and Marco-Pallarés, 2016). Additionally, theta oscillations are sensitive to reward probability, updating and enhancing the predictive power of individuals towards more optimal behaviours (Cohen and Cavanagh, 2011). These findings land the hypothesis that both the FRN and theta activity may be used to measure and understand reward prediction errors and anticipatory behavioural responses (Marco-Pallarés et al., 2008).

While some studies have found a link between theta activity and reward magnitude (Hajihosseini et al., 2012; Gheza et al., 2018), others have not (Marco-Pallarés et al., 2008). Additionally, some research suggests that theta activity is insensitive to performance assessment and is instead influenced by breaches of reward expectation (Cavanagh et al., 2011; Mas-Herrero and Marco-Pallarés, 2014), but others find that no such link exists between theta and reward expectation (Watts et al., 2017). For instance, in research conducted by Mas-Herrero and Pallarés (2014), a one-step decision task was proposed where, in each trial, participants were compelled to choose between two geometric figures. Participants were required to determine the most favourable figure using trial-and-error feedback. The most rewarding figure changed in each block of approximately 20 trials without knowledge from participants. The results evidenced the association of theta activity with RPE both in positive and negative feedbacks, suggesting a role of this oscillatory response in the computation of the absolute value of reward prediction error (surprise, Mas-Herrero and Marco-Pallarés, 2014).

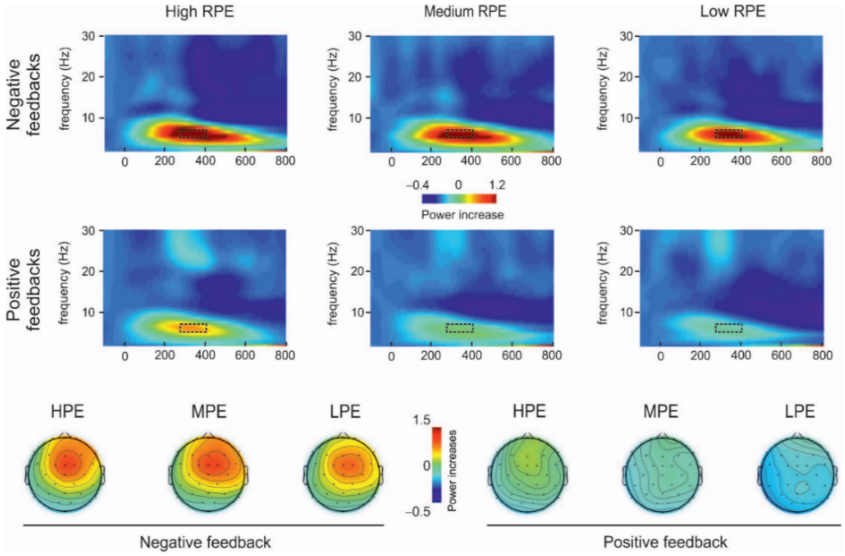


Figure 1.5. Time-frequency power changes for negative and positive feedback in trials with three different reward prediction error magnitudes: high (left), medium (mid) and low (right). Note that in both positive and negative feedback, the theta activity increases with RPE. Reproduced from Más-Herrero and Marco-Pallarés (2014) with permission.

Beta and theta band oscillations respond differently to feedback stimulus characteristics. Andreaou et al. (2017) employed single-trial coupling of concurrent fMRI and EEG data to analyse networks associated with oscillatory responses to feedback. They detected distinct connections between theta and high-beta oscillations and non-overlapping brain networks. In particular, increased beta power in response to positive feedback was related to activation of the core reward network regions. In comparison, a rise in theta-band power following a loss was related to the activation of a frontoparietal network that included the ACC. These findings imply that positive and negative feedback might be partially related to distinct brain networks and that, within these networks, the communication is mediated by distinct oscillatory activity. Alternatively, these results could suggest different functional roles for these two oscillatory components, being

theta activity associated with the computation of (absolute value of) reward prediction errors (Mas-Herrero and Marco-Pallarés, 2014, 2016) and beta-gamma activity related to highly relevant or unexpected positive events (Marco-Pallarés et al. 2015).

Hierarchical Reinforcement Learning (HRL) and the problem of dimensionality

The RL models have impacted various disciplines such as neuroscience and computational science, being a widely extended model with exuberant empirical validity. Although they theoretically allow behaviour to be tracked over time, very often, this seems to be the case only in simple behavioural schemes. Therefore, despite their descriptive capacity, RL models have been confined to experimental designs that fall short when describing complex situations (Hengst, 2012). Sometimes called the curse of dimensionality, RL models lack the ability to operationalise routines of actions that require multiple choices (Botvinick, 2012; Eickstein and Collins, 2020). Thus, computational efficiency deteriorates as the size of the learning problem increases and the chains of actions necessary to achieve a goal extend. This problem of scale has obvious relevance for the applications of RL within the DMN since daily life learning problems are notoriously larger and more complex than the typical RL experimental designs since multiple environmental states and action-reward relationships are involved (Botvinick, 2011).

A proposed solution to overcome the limitations of RL models in describing more realistic situations might come from the hierarchical organisation of behaviour. For instance, to reduce complexity and

uncertainty, individuals set a goal and divide it into sequences of steps, with all the actions necessary to achieve the final goal. Therefore, this set of individualised actions would be hierarchically organised to achieve a reward; thus, an action could be subordinate and guided towards a higher ordinal level and closer to an expected result. In fact, actions are organised into chains of sub-tasks that fit together to achieve a general goal. This is how hierarchical reinforcement learning (HRL) is introduced, where the functionality of the standard RL model is modified to incorporate pre-states necessary to access a final reward (Hengst, 2012). Therefore, HRL refers to a computational model that aims to describe temporally more abstract learning processes and behaviours involving sub-tasks and sub-objectives.

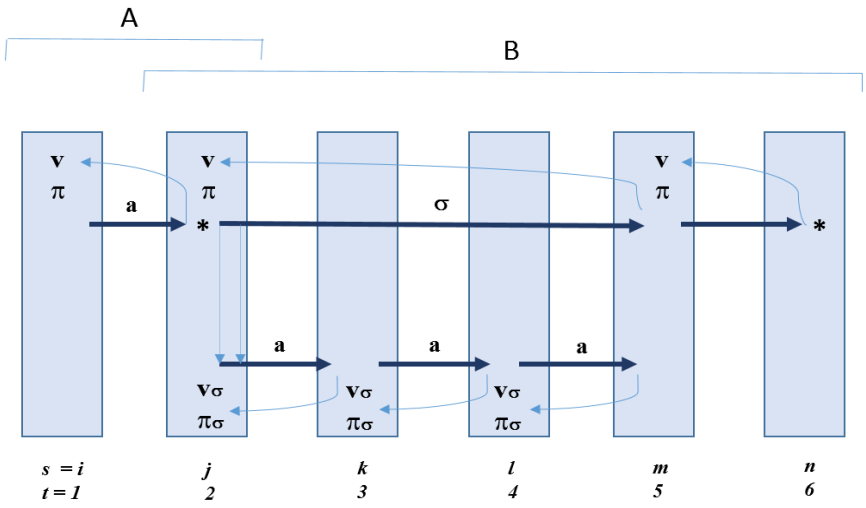


Figure 1.6. Relationship between the standard TD model and the HRL framework. The agent meets a new state and picks an action in each section. Each state is associated with a value (v), and each action is associated with a policy (pi), which the agent uses to forecast future states and actions. The agent generates a reward prediction error after reaching the resulting state (s=i) and collecting the primary reward (last box). Reproduced from Botvinick (2012) with permission.

HRL algorithms seek to solve this scalability challenge by grouping states and actions to build higher-level behavioural plans. These ordinal enchained

sequences consist of organised single actions aimed at specific sub-goals. As a consequence, options can be learnt even if there are middle states (pre-states) towards the ultimate objective. Thus, a shorter concatenate temporal resolution of such pre-states results in decreasing the task complexity and resolving the dimensional problem of extensive behavioural schemes (Figure 1.6). Importantly in HRL models, learning takes place by simple extensions of conventional RL: action selection that results in better than expected outcomes is reinforced, whilst successful completion of a selected option acts as a pseudo-reward that reinforces previous lower-level actions based on the same RL principles. The learning challenge is addressed at several levels of abstraction at the same time, discovering both low-level actions and high-level options that most effectively achieve their respective goal or sub-goal.

Therefore, while a conventional RL agent chooses between concrete and straightforward actions, the HRL agent requires the implementation of sub-routines, each of which has its own characteristics and associated policies, which are used to achieve sub-goals. Therefore, TD mechanisms and RPEs encoding allow the agent to learn which subroutines could be appropriate for each state and sub-goal (Hengst, 2012). By including temporal abstraction in each subroutine, the dimension or scalability problem described above is greatly alleviated.

Additionally, this MDP installed in each sub-routine allows that upon reaching a sub-goal, the learning is updated. Notably, unlike top-level objectives, relevant sub-goals are generally not associated with primary rewards. This complicates the challenge of sub-goal setting and its incidence in TD mechanisms, which is probably the most challenging in HRL (Botvinick, 2012; Wiering and van Otterlo, 2012). Beyond the definition of

the sub-goals, there are expected results for each sub-task, and therefore prediction errors are generated, known as pseudo reward prediction errors (PRPE). Presumably, as in standard RL, goal achievement has a motivational effect, and this is potentially the case with sub-goals. Accordingly, as with rewards, obtaining these pseudo-rewards by completing sub-tasks might guide learning and influence decision-making.

Neural networks of sub-goals and HRL

HRL was designed to outperform typical RL techniques in computing efficiency by describing learning during long sequences of actions, offering a more naturalistic approach. This is particularly relevant when describing realistic day-to-day life actions, which require a higher level of complexity rather than a single choice. HRL contains the notion of consecutive choices, which describe action plans composed of sequences of simple actions (Chiang and Wallis, 2018; Botvinick et al., 2009; Holroyd and Yeung, 2012). As an example, a set of *actions* might consist of the individual steps required to drive a bike (unleashing a padlock, taking it out of the parking lot, pedalling for ten minutes, and so on). In contrast, an *option* might consist of the sequence of actions that leads the biker to a specific park in the town.

Importantly, each choice and policy carry the specified sequence of activities and the complete collection of steps that chart different alternatives to reach the target state. As it happens in standard RL, choices are defined by their association with goals and the available information of the states. In RL, the stimuli salience and state policies map the transitions from initiation to goal states. Moreover, HRL enables option levels (such as going to

another park) and single-action levels (such as taking the Diagonal Avenue and stopping by a friend's house to share a coffee).

Botvinick and colleagues (2011) explored which neural networks could be involved in HRL. Particularly, the effort of bringing together cognitive neuroscience and HRL started by fitting the concept. Both cognitive neuroscience and computational disciplines aim to discover the set of actions between stimuli and responses, intrinsically assuming a task-dependent behaviour (Niv, 2009). While early findings suggested a predominant role of the prefrontal cortex in selecting options but also in monitoring actions, more recent developments defend a division between agent and critic with their respective neural correlates (Holroyd and Yeung, 2012; Pezzulo et al., 2018).

Elaborating on the idea of learning impacting decision making, and particularly from a model-based (goal-oriented) behaviour, there are different systems running in parallel to encompass the whole reward circuit dynamics. Indeed, in a more suitable approach to the idea of the actor-critic in model-based learning and diving into the neural representation of learning through contingencies, the dorsolateral prefrontal cortex (DLPFC) and motor-related components in the dorsal striatum (which comprise the actor) execute the different options, and the orbitofrontal cortex and the ventral striatum (which comprise the critic) evaluate progress toward the options' goal states (O'Doherty et al., 2004).

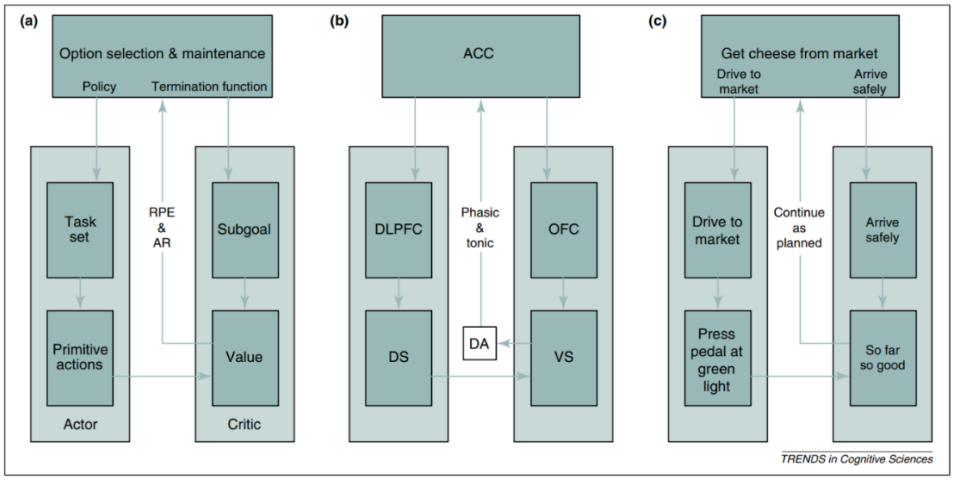


Figure 1.7. Relationship between ordinary actions (*Section c*), Hierarchical Reinforcement Learning framework (S*ection a),* and Neural correlates (*Section c*), following the actor-critic architecture. In *Section b*, the dorsal striatum implements the actor, whereas the ventral tegmental region (VTA) and the ventral striatum (VS) collaborate to implement the critic. Reproduced from Holroyd and Yeung (2012) with permission.

Therefore, under this architecture (see Figure 1.7), the dorsal striatum carries out the actor's policy while the ventral striatum, which is at the main substrate of the critic, monitors and weights current events' accuracy to anticipate future rewards (Van der meer and Redish, 2010; Cohen and Frank, 2009). In addition, DLPFC sends top-down biasing signals to the dorsal striatum that update the current policy, and the orbitofrontal cortex provides information about higher hierarchical goals to the ventral striatum (Caligiore et al., 2019; Houk et al., 1995; O'Doherty et al., 2004). Finally, although there are different proposals on its role, ACC would be mainly involved in action selection (Holroyd et al., 2016; 2017) or would play a more integrative approach (Diuk et al., 2013b).

Another important aspect when trying to uncover the brain networks involved in HRL is the role of sub-goals. The theory of HRL is based on the

claim that while learning to attain a goal through a sequence of actions, therefore agents chunk into sub-tasks and evaluate if an alternative meets a particular sub-goal. (Behrens and Jocham, 2011). HRL's important innovation is divided into two components. First the assumption that sub-goals main purpose is fulfilling an overarching task goal, which requires a hierarchical representation of a foreseen sequence of actions. Second is the assumption that all sub-goals serve as previous states toward the ultimate or overall goal.

Achieving a sub-goal would result in attaining a pseudo-reward. Nevertheless, it is central to differentiate pseudo rewards as previous and different to the final primary reward. In contrast to the Reward Prediction Errors (RPE) elicited by primary or secondary rewards, which indicate the encoding of such interaction between agent actions and a consequence (achieving or not a final goal), pseudo-reward prediction errors (PRPEs) in HRL are distinct and indicate the various levels of a pertinent sub-goal and its association to specific pseudo-rewards (Botvinick et al., 2009). As reviewed above, previous research has delineated the role of diverse brain areas such as the Anterior Cingulate Cortex, Pre-Frontal Cortex and Ventral Striatum in the experience of learning, particularly within the scheme of RL (Niv, 2009). Furthermore, during the last decade, some studies have supported the relevance of pseudo-rewards in guiding goal-directed behaviour. The advent of a more naturalistic and efficient approach in HRL also raised questions about the neural correlates underlying the encoding of PRPEs.

In this line, Botvinick et al. (2009) assessed the parallelism between aspects of the actor-critic architecture for RL and certain brain areas previously connected to the brain activity during RL experimental paradigms

(Sutton and Barto 1998). Botvinick et al. (2009) investigated what additions or changes had to be made to extend the actor-critic assumptions to suit HRL conventions. Some research proposed that just a few changes were required, and each of them was consistent with proven previous neuroscientific discoveries (Botvinick et al. 2009).

In this study, an important computational requirement of HRL was the maintenance of representation of the currently selected choice. In other words, every state operates individually as a single RL standard problem, but it is linked to overarching hierarchies. This particularity suggested the involvement of the DLPFC and other frontal regions, such as the pre-supplementary motor area (pre-SMA). Furthermore, neurons in different frontal regions have been found to signal for specific sequential actions in the same way as alternatives do in HRL. Prior findings also suggest that frontal cortical regions signal actions at several and nested levels of temporal assembly (Diuk et al., 2013a; Badre, 2008; Koechlin et al., 2003), similar to how HRL representations arrange activities into hierarchies, with policies for one choice and one state.

Ribas-Fernandes et al. (2011) tested the presence of specific neural correlates of PRPE by using a video game named Courier Task (Figure 1.8). This task's primary purpose was to expeditiously perform a "package delivery". Participants were told that there would be three stimuli (truck, box, and house) on the screen and were instructed to use the joystick to steer the vehicle to first pick up the package and then deliver it to the house. The relationship between these three locations in the courier task provided a hierarchical structure, where the delivery was the overarching goal. The main manipulation was called *the package's jump event*, which caused the package to travel to a random, unexpected position before the trunk could reach it.

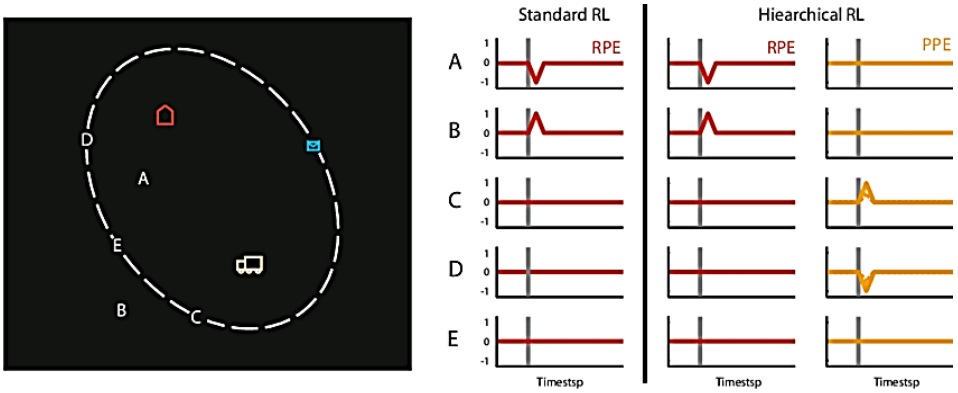


Figure 1.8. The design of the Courier task is displayed in the left panel. In both reinforcement learning and hierarchical reinforcement learning, the right panel displays the instances in which reward and pseudo-reward prediction errors are produced. Before a jump event happens, grey bars show the preceding time step. Reproduced from Ribas-Fernandes (2011) with permission.

According to RL, a jump event may induce a positive or negative RPE based on the decrease and increase in the overall distance required to deliver the box. In addition, HRL implies that a jump event may induce positive or negative PRPEs, in this case, depending on how close the truck is in reference to the new location. On the basis of these hypotheses, converging findings from EEG and fMRI experiments revealed that a jump event that increased a sub-goal distance without affecting the total distance evoked a negative PRPE and that the features of this neural signal were formed in the anterior cingulate cortex (ACC) (Ribas-Fernandes et al., 2011; 2019).

These findings, and particularly the contribution of the mPFC, have been tested and validated later on. For example, Ribas-Fernandes et al. (2011) focussed on sub-goals with negative outcomes, but in a further experiment, Ribas-Fernandes et al., 2018 showed that ACC was also involved in positive PRPE (Ribas-Fernandes et al., 2018). In another study with primates, Chiang and Wallis (2018) trained two rhesus monkeys to complete an adapted version of the delivery task. They measured the electrical activity of neurons

in the ACC, lateral prefrontal cortex, and orbitofrontal cortex during this task. Their results showed that a small group of neurons in the ACC encoded PRPEs, supporting the key role of this area in HRL. Therefore, these studies support the critical role of ACC in HRL.

However, given the importance of the reward network in RL processing, it is reasonable to suppose that areas of this circuit should also be involved in HRL. In order to investigate this topic, Diuk et al.(2013b) employed functional neuroimaging to calculate prediction error signals in people undertaking a hierarchical task that required simultaneous encoding of not just a global-unitary RPE but PRPEs. They designed a task with two levels, where intermediate states were needed to meet the final goal of winning a casino (Figure 1.9).

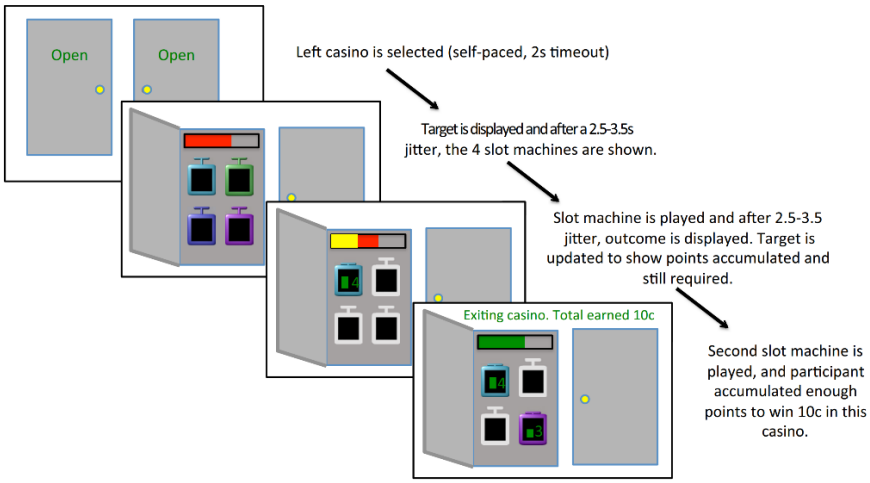


Figure 1.9. The task design of Diuk et al. 2013b. It involves picking between options at two distinct levels. In the initial phase of each trial, the participant must select one of two casinos. The selected casino then displays four different-coloured slot machines. Next, the participant must select one of the four machines displayed; for instance, the participant plays the slot machine in the upper left corner, and the points gained on that machine are indicated by a green bar. The red bar represents the expected amount of points, while the yellow bar represents the actual number of points earned**.** Reproduced from Diuk et al. (2013b) with permission

This task was created to induce learning on two levels.: first, the level where participants selected a casino and then the level where they selected a slot machine. After playing the slot machine (whichever was chosen), two unique and coincidental prediction errors should occur when the outcome of that machine was presented concurrently with the casino's overall win/loss conclusion. When compared to several other feasible alternative models, they found that the HRL model best described the participants' behaviour and that both RPEs and PRPEs activated VS. These findings have two main ramifications: The first is that, as needed by HRL, the human brain can calculate prediction errors that span several states and actions in time (Botvinick et al. 2009). In contrast to prior research, which was focused on the computation of a single prediction error signal, the second conclusion is that numerous prediction error signals may be formed and employed for learning in the brain (Schultz et al., 1997).

In recent research, Mas-Herrero et al. (2019) studied if pseudo-rewards could bias decision behaviour *in situations in which obtaining more pseudo-rewards did not imply a higher final reward* and whether this bias could represent striatal sensitivity to pseudo-rewards. To test these predictions, they devised an fMRI learning task in which participants were needed to, in the first place, complete a sub-goal (opening a padlock) to get a probabilistic monetary reward. Participants demonstrated a considerable inclination for the key that unlocked more boxes (i.e., the box that delivered more pseudo-rewards), despite the fact that it did not result in an increase in monetary reward. Then, fMRI data revealed a parallel ventral striatum activation not just for the reward but also for the pseudo-reward prediction errors encoding (Figure 1.10). Additionally, individual variations in behavioural preference for the most pseudo-rewarding option were predicted

by the striatal sensitivity to the prediction errors coming from processing the pseudo rewards.

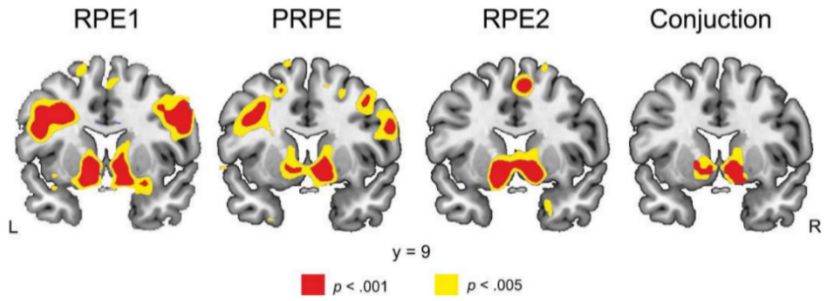


Figure 1.10. Results of the contrast for the different prediction errors in Mas-Herrero et al. 2019. They demonstrated that both RPEs and PRPEs are encoded concurrently in the ventral striatum. This conclusion is justified on the grounds that the striatum is required for computing prediction errors at multiple hierarchical levels. Reproduced from Mas-Herrero et al. (2019) with permission.

Importantly, and as stated above, the findings by Mas-Herrero et al. (2019) demonstrate a bias in participants' decisions, which is induced by a preference for pseudo-rewards. However, this bias cannot be considered sub-optimal because both alternatives were rewarded exactly in the same proportion, and hence the bias for one alternative was not linked with a cost, meaning fewer rewards if there would have been a bias. But, would the same preference for pseudo-rewards continue even at the cost of the final reward leading to a sub-optimal decision? This will be one of the key questions of the present thesis.

Sub-optimal Decision Making

Regardless of the capacity of pseudo rewards to guide behaviour in a goal-directed setting, it is unclear if the individual's engagement toward pseudo-feedback acquisition could derive from a sub-optimal pattern. To understand the potential incidence of pseudo-rewards in sub-optimal decision-making, it is crucial to explore the idea of sub-optimal choice behaviour and maladaptive behaviours.

For instance, conventional approaches of Decision Theory, such as the foraging theory (Stephen and Krebs, 2019) and rational choice theory (Scott, 2000), propose that a person should pick choices that maximise reinforcement with the least effort possible. Therefore, any learning of the information present in the environment would only contribute to building up an optimal decision. Indeed, any significant deviance from the most favourable alternative is an indicator of sub-optimal choice behaviours, also known as maladaptive decision-making (Zentall and Stagner, 2011; Swintosky et al., 2021). Sub-optimal choice behaviour has been studied in human pathologies such as gambling, substance abuse, and eating disorders (Bechara and Damasio, 2002; Brogan et al., 2010). For example, in gambling disorders, research has provided large evidence of risk-taking behaviour underlying sub-optimal choices (Swintosky et al., 2021).

Despite the negative consequences associated with the choices made, maladaptive decision-making can be persistent and recurrent, limiting the individual's chances to exploit an available and more optimal option. Importantly, some traditional theories failed to explain behaviour that was far from rational decisions. To explain the nature of such behaviour, different accounts proposed that they could be related to various factors such as the

differential impact of the salience of wins vs losses (Tversky and Kahneman, 1985) or misperception of probability (Sanbonmatsu et al., 1997). However, beyond pathology and even given these factors, there is still debate on the factors underlying some sub-optimal decision-making behaviours and their underlying neural mechanisms.

As discussed before, individuals make decisions by weighing the value of options available to choose between them. Certain options, such as whether to get more or less reward, are straightforward, and optimal decision-makers should pick the best option, but these decisions might be more difficult when more options with different properties (e.g., different probabilities of reward, magnitudes, expected values) are available. In any case, individuals reckon for a trade-off between risk and increased reward (Smith et al., 2017).

However, these decisions are also highly influenced by individual differences. Therefore, animal models have been useful to better identify the processes underlying sub-optimal choice behaviour (Zentall, 2016; Zentall and Case, 2018). In this regard, one commonly used experimental scheme that has been successful in this quest is to give animals a choice between two options, each of which leads to a distinct stimulus reward, with one option being sub-optimal in comparison to the other, and then to examine the decisions made (Figure 1.11).

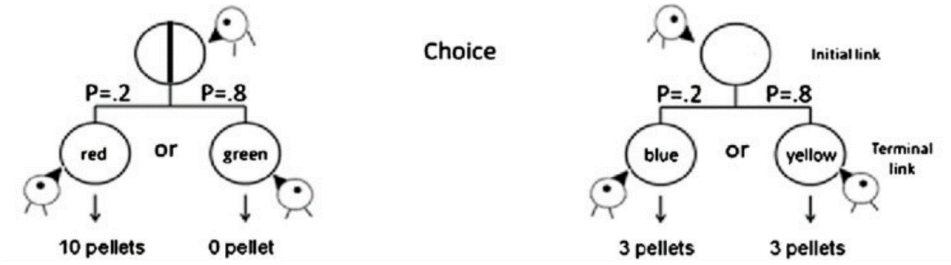


Figure 1.11. The graph shows a paradigm in which pigeons choose between two options. One alternative provides colour with p=0.8, but then pellets are not provided, and another colour with p=0.2 and then ten pellets are delivered. The other alternative presents different colours with the same probability as the previous alternative, but in this case, the two colours deliver three pellets. Pigeons prefer the first alternative, even when this choice is sub-optimal. Reproduced from Zentall et al. (2011) with permission.

Using such a technique, McDevitt et al. (2016) and Zentall et al. (2011) have shown that, under certain conditions, pigeons exhibit a sub-optimal preference for the option that provides less reinforcement over an alternative that provides more reinforcement. Indeed, such behavioural bias might signal cues that indicate the availability of reinforcement without altering the overall likelihood of reward, therefore without modifying their evaluation of final rewards. For instance, pigeons will step on a treadle solely to register a discriminative stimulus indicating the effect on action. In other words, seeing outcomes alter the RL scheme. In fact, pigeons prefer an alternative that generates discriminative stimuli (i.e., a light linked with 100% reinforcement) over an option that generates stimuli related to a probability of either being reinforced or not, the two alternatives being reinforcement (Stagner and Zentall, 2010). Furthermore, the effects exhibited by the sub-optimal choice technique in pigeons have been replicated in human participants (Molet et al., 2012; McDevitt et al., 2019), suggesting that decision-making might be influenced by maladaptive behaviours. This

deviation from optimality has prompted questions about how choice behaviour has evolved in relation to the decision-making mechanisms (Monteiro and Kacelnik, 2015; Swintosky et al., 2021).

Another proposal is that feedback-predictive cues and, in general, reward-related informative cues have a disproportionate effect on the decision in the sub-optimal choice because they give information about availability, and this amount of information acts as a conditioned reinforcer (Zentall, 2016). The difficulty in clarifying sub-optimal choice with regard to conditioned reinforcement is determining how these stimuli function act as conditioned reinforcers and the reason why the sub-optimal option's conditioned reinforcer would be favoured above the ideal alternative. (Vonder Haar, 2019; Cunningham, 2020).

Different accounts have tried to explain these behaviours. The *Signals for Good News* hypothesis (McDevitt et al., 2016) proposes that, similarly to the delay discounting, the value of a conditioned reinforcer is proportional to the delay of cue signals. Therefore, if a cue suggests the presence of a reward, it can bias choice even when it is not leading to a greater benefit than the other options with less signalling or cues. As in the observing behaviours described in pigeons, McDevitt (2016) hypothesised that anytime reward-related information appears, it tends to generate a preference over those alternatives that bring uncertainty and do not always yield to a reward, even when probabilistically they are more beneficial. Another approach suggested by Beierholm and Dayan (2010) claims that sub-optimal choice behaviour might come from disengagement during reinforcement. Therefore, the sub-optimal choice would occur when animals disengage or stop paying attention to lower-valued conditioned stimuli that are actually leading to a larger outcome. According to this proposal, when there are different

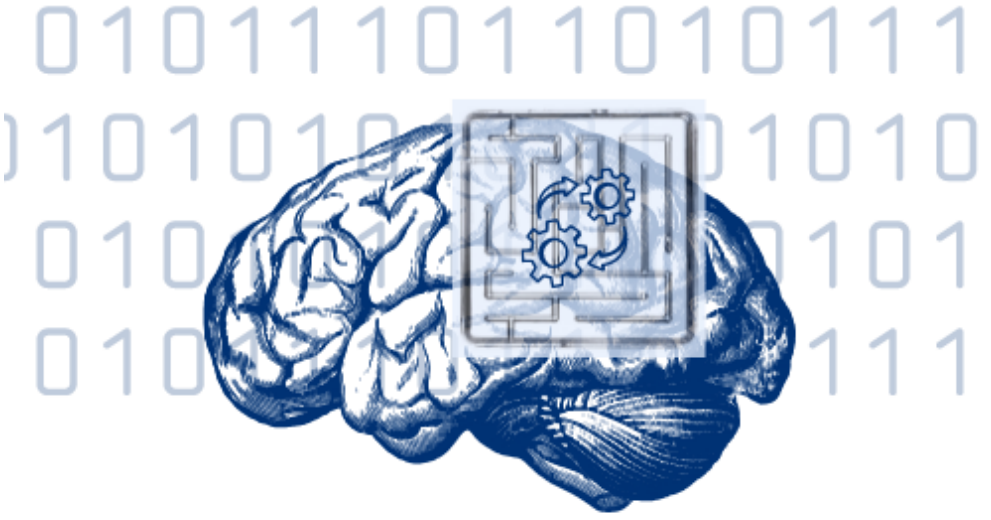
available options, but there is more valuable predictive information about the alternative that happens to be less optimal, individuals might engage and exploit it. Therefore, there would be a forgetting ratio on the feedback processing coming from the options that provide less information linked to a final reward, even when they are overall better options. In consequence, when disengagement to lower-valued cues occurs, the individuals decimate the chance of learning from the reinforcer. In the described situation, one cue is more predictive of reward; at the same time, one option is more predictive of non-reward but does not necessarily provide fewer benefits. Consequently, individuals undervalue feedback processing coming from those cues with less predictive value since they are unable to correlate any output and its correspondent predictive signal (Beierholm and Dayan, 2010). In this scenario, individuals do not disengage from the cue related to the most positive reward (more predictive of reward), even if it is not the best available overall option. Hence, individuals acquire a bias of the overall probabilities, ignoring the learning from those outcomes coming after cues less related to reward (McDevitt et al., 2016). As a result, individuals may demonstrate a preference for the most signalled option, even when it is less advantageous.

Finally, interesting differences appear when comparing habitual and goal-directed behaviour. Therefore, habitual processes are frequently viewed as the underlying mechanisms of a variety of sub-optimal behaviours (Decker et al., 2016). Therefore, if habitual behaviours are those mostly driven by the stimulus salience and goal-directed behaviour are those oriented to a specific objective, the former might be more prone to be associated with the choice of the least beneficial option due to the force of the habit. Following this idea of two competing decision-making systems, the values determined using model-based RL for goal-directed behaviours

and model-free RL for habitual behaviour are both reasonable yet not always identical (Niv, 2019). Indeed, given that these systems compute action values using distinct algorithms, they are each uniquely sensitive to situational factors, such as cues and feedback-related information, that might result in inefficient or seemingly sub-optimal behaviours (Decker et al., 2016; O'Doherty et al., 2017).

In the case of habitual behaviours, the brain uses both learning strategies (*stimulus-driven and goal-oriented*). Despite this, the cognitive demands of such decision-making need the exclusive use of limited resources, precluding the performance of several tasks. Besides, habitual behaviours and model-free reinforcement learning mechanisms are always ready to preserve the behavioural strategy of multitasking. Therefore, the two-level tasks typically used to explain hierarchical learning might be the tip of the iceberg to understanding sub-optimal behaviour as a transition from habitual behaviour to goal-directed behaviour (Decker et al., 2016).

To summarise the information presented so far, our daily actions are distributed in sequences in order to achieve objectives. We divide and organise our tasks in steps prior to a final goal, which in turn defines sub-goals. We hierarchise in the process of simplifying our decision-making. Previous studies have also provided evidence of HRL and its correlates, yet under the pivotal assumption that behaviour is progressively getting refined to be more optimal, but, at the same time, they have shown that people choose options with more pseudo rewards due to sub-goal attainment when different options lead to the same final reward. But, could the engaging power of pseudo-feedbacks bring individuals into sub-optimal choice patterns? And if so, what would be the underlying neural mechanisms of such sub-optimal behaviours?



Chapter 2. Research Aims

,

Chapter 2: Research aims

The current doctoral thesis' main objective is to **investigate the behavioural and neural correlates of pseudo-feedback processing in sub-optimal decisions**. Given previous studies showing the preference of people for options that yield more pseudo-rewards, we hypothesise that the accomplishment of sub-goals might bias decisions towards sub-optimal behaviours and that the brain mechanisms underlying such bias will be those involved in the computation of reward and pseudo-reward prediction errors. Therefore, the specific objectives of the present thesis are:

Specific Objective 1. ***To examine whether the preference for pseudo-rewards might lead to sub-optimal decisions*****.**

Previously, tasks with secondary goals have been used to minimise overall task complexity (Ribas-Fernandes et al., 2011; Diuk et al., 2013b; Mas-Herrero et al., 2019), finding that intermediate phases have an impact on learning. Therefore, sub-goals drive learning while offering stronger predictive values for pre-states that anticipated a final reward. Furthermore, these pseudo-feedback prediction errors engagement might induce sub-optimal choice behaviours as individuals guide their behaviours from one sub-goal to the next one due to accomplish an over-arching goal. We hypothesise that the current information on intermediate processes may lead to poorer decision-making as a result of learning from pseudo-feedback, leading to sub-optimal behaviour. We will develop this goal in all the studies, but particularly in Study 1, which will contain two different experiments. The experiment 1 will act as a stress test of a previously reported paradigm (Mas-Herrero et al., 2019); therefore, different versions of a two-step

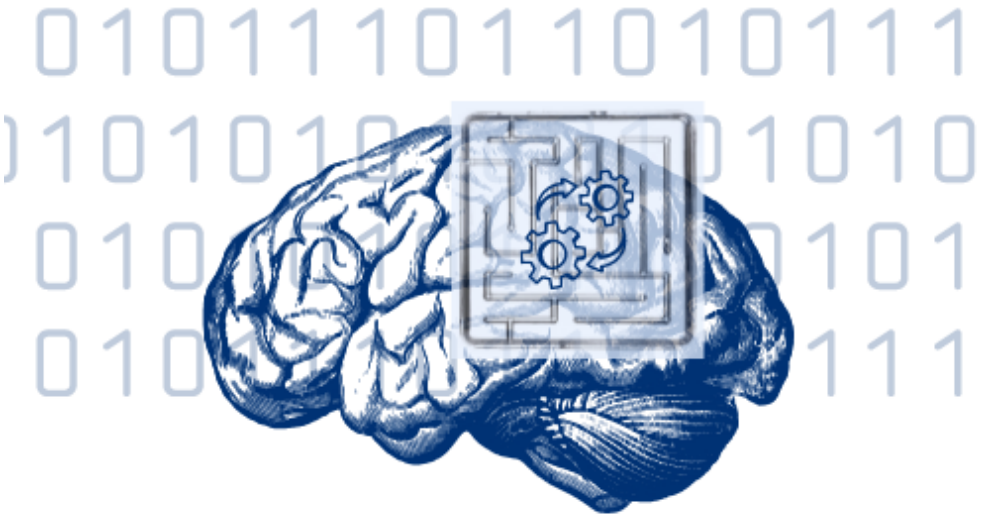
hierarchical task will be displayed. The likelihood of receiving that pseudo-feedback will be tested to validate whether or not participants show the preference for the most pseudo-rewarding option when the probability of accessing pseudo-feedback is more or less evident. The experiment 2 will use a mobile probability trying to determine the point of subjective equivalence (PSE), at which individuals make a trade-off between pursuing pseudo-reward and maximising the ultimate reward. We hypothesise that participants will prefer the most pseudo-rewarding option even when it leads to lower final rewards.

Specific Objective 2. ***To explore the computational and brain oscillatory mechanisms involved in sub-optimal behaviour driven by pseudo-rewards***.

To reach this goal, we will perform an EEG experiment to describe the brain oscillations underlying the pseudo-reward-prediction error (PRPE). We hypothesise that theta oscillatory activity has been associated with activation of the anterior cingulate cortex (ACC) and encoding of prediction errors (Holroyd and Yeung, 2012; Gruber et al., 2013; Mas-Herrero and Pallarés, 2016; Shahnazian et al., 2018), will also be involved in the computation of pseudo-reward prediction errors and in the bias towards sub-optimal behaviours driven by the preference for sub-goals. Importantly, in order to study reward and pseudo-reward prediction errors, we will develop an RL computational model to account for the complexity of the task, and that should be able to demonstrate the parallel learning at both action levels of the two-step task.

Specific Objective 3. ***To explore the brain regions involved in pseudo-feedback processing under sub-optimal behaviour and its relationship with a Hierarchical Reinforcement Learning model.***

It is well known that striatum-projected dopaminergic neurons play a critical role in the Reward Prediction Errors computation (Schultz et al., 1998). Furthermore, given that value functions can vary because of reward attainment, as well as pseudo-feedback processing, we hypothesise that both RPE and PRPE will present similar ventral striatum contributions, probably suggesting that the brain encodes them in parallel in order to reach not only the history of rewards but also to the pseudo-feedback encoding (Ribas-Fernandes et al., 2011; Diuk et al., 2013b; Mas-Herrero et al., 2019). To test this hypothesis, we will use the two-step task in an fMRI setting and explore the brain correlates of RPE and PRPE computed using an RL computational model.



Chapter 3. Study 1

Chapter 3. Study 1. Sub-optimal Choice Behaviour is driven by pseudo-reward sub-goal attainment

Summary

Our daily actions are distributed in sequences in order to achieve objectives. We divide and organise our tasks in steps prior to a final goal, which in turn defines sub-goals. We hierarchise in the process of simplifying our decision-making. Previous findings have shown that people choose options with more pseudo rewards due to sub-goal attainment. In the present study, 226 university students participated in two experiments to test the hypothesis that pseudo rewards bias decisions. The task consisted of two alternatives, presented as two shapes in two steps. The results of this experiment revealed that people preferred the more pseudo-rewarding option, assuming the cost of a lower final reward. This study demonstrates that people consistently prefer pseudo-rewards, which affects their decisions and leads to sub-optimal behaviour. These findings open up possibilities in the study of pseudo-reward stimuli and their influence on behaviour.

Introduction

In our daily lives, we constantly make decisions based on anticipating and receiving rewards. Often, achieving a goal requires delaying gratification in favour of decisions that will bring agents closer to the ultimate desired outcome. These extended behavioural strategies are essential for reducing uncertainty in a temporal frame, and they require breaking down tasks into

progressive steps whose completion signifies getting closer to a reward. Agents use sub-goals to determine the correctness of their actions prior to reaching the end of a sequence when the final objective is distant in time or requires multiple decisions. The division of the task into a series of steps with their corresponding feedback regarding the correctness of the actions (pseudo-feedbacks, Botvinick, 2012; Mas-Herrero et al., 2019) decreases the uncertainty of the overall chained action and may decrease biased behaviours (Brandtstädter and Rothermund, 2002).

In fact, the information received after the correctness or incorrectness of each intermediate step (pseudo-feedback) is beneficial for accelerating the learning process. Previous research has shown that the processing of prediction errors in pseudo-feedbacks engages similar neural mechanisms as final feedback (Ribas-Fernandes et al., 2018; Mas-Herrero et al., 2019) and that the achievement of such pre-states and informative cues related to the reward can be intrinsically rewarding (Clark and Gilchrist, 2018). On the other hand, numerous studies have demonstrated that cues signalling the presence of unconditional stimuli (e.g., food) may attract the attention of animals, thereby becoming motivational and resulting in maladaptive behaviour.

For instance, pigeons may peak a key light that has been paired with food delivery, even if this action has no effect or if this action prevents them from receiving the food (Hearst and Jenkin, 1974; Dorfman and Gershman, 2019). While this sign-tracking phenomenon has been extensively described in animal models, recent research has revealed a similar pattern of behaviour in humans who have undergone Pavlovian conditioning (Garofalo and Pellegrino, 2013, Colaizzi et al., 2019).

Therefore, it appears that signals indicating the possibility of a reward are reinforcing themselves (either because they have been paired with them or because they inform about the correctness of a necessary action to attain them). Despite the fact that available information may improve behaviour, it may also lead to maladaptive behaviours. Numerous studies have demonstrated that animals exhibit observing behaviour, i.e., a preference for richer sources of information that can predict rewards while failing to accurately encode other alternatives (Dayan and Beirholm, 2010). This attempt to find a trade-off between information seeking and optimal decision-making may result in sub-optimal behaviour due to the rewarding properties of cues' utility (Fu and Gray, 2006; Wickens and McCarley, 2019). Therefore, agents may decide to engage in options that provide more information, even if doing so incurs costs (Eliaz and Schotter, 2010; Morrison et al., 2015; Zentall, 2016). However, it is still unknown to what extent pseudo-reward feedback could act similarly, i.e., to what extent receiving positive feedback regarding the correctness of an intermediate action could bias performance towards the acquisition of positive pseudo-feedbacks rather than final rewards.

In the present study, we will investigate this question by modifying an earlier two-step paradigm. In the original version of this paradigm (Mas-Herrero et al., 2019), participants were required to choose between two keys that could or could not unlock a padlock, which could yield a reward. One key was more likely to open the padlock, but it delivered the reward less frequently, whereas the other key opened the lock less frequently but delivered the reward more frequently. This experiment demonstrated that participants preferred the key that opened the padlock more frequently, despite the fact that both options had an equal chance of receiving a final reward. As a result, the event of unlocking the padlock (pseudo-feedback)

proved to be sufficiently rewarding to create an association between this particular key and the ultimate objective. In fact, this study was able to demonstrate a preference for the pseudo-rewarding option when there was no cost associated with the final reward; that is, this preference did not result in fewer benefits. Therefore, the hypothesis that positive pseudo-feedbacks may lead to sub-optimal decision patterns has not yet been tested.

The purpose of this study was to investigate the preference for pseudo-rewards in two separate experiments. In the first experiment, we attempted to replicate and extend the findings of Mas-Herrero et al. (2019), specifically regarding the preference for the most pseudo-rewarding options in similar experimental designs. In the second experiment, we manipulated the probabilities of the selected options in order to examine this preference for pseudo-feedbacks, with the less preferred option presenting a greater probability of final reward. We hypothesised that participants would show a clear preference for the most pseudo-rewarding option when the two options had the same expected value of the final reward (experiment 1) and that they would exhibit sub-optimal behaviour toward the most pseudo-rewarding option even when it yielded fewer final rewards (experiment 2).

Methods

Experiment 1.

Participants.

One hundred eight healthy students from the Faculty of Psychology of the University of Barcelona (M = 21.5 years, SD = 4.1, 25 men) participated for course points in the experiment. Participants were divided into three groups

of 36 individuals each, corresponding to different task versions. All participants gave written informed consent, and the University of Barcelona's Ethical Committee approved all procedures.

Experimental procedure

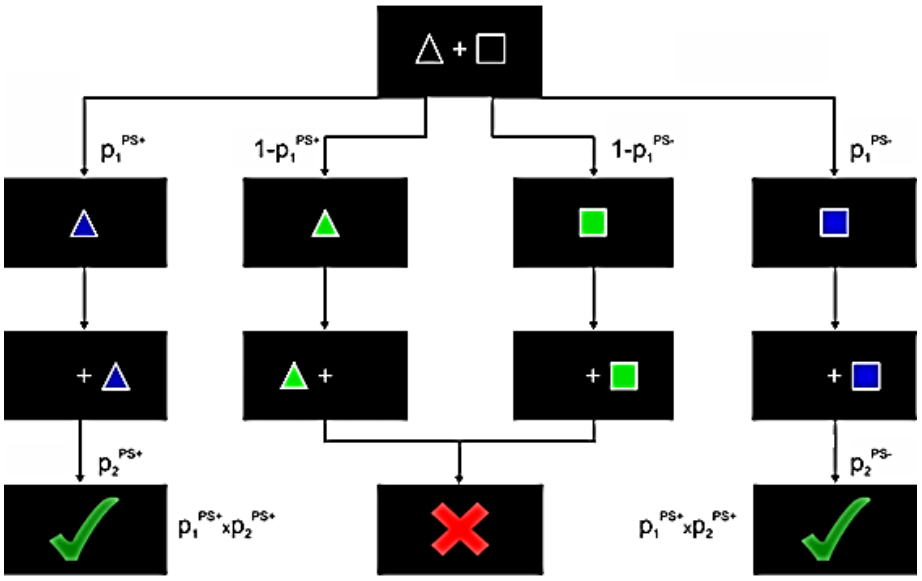


Figure 3.1. Task design. Two figures were associated with different probabilities of obtaining pseudo-rewards (coloured figures; p1) and, after them, obtaining the final reward (green tick, p2). PS+ refers the option that delivers pseudo-feedback more often, while PS- indicates the option that delivers pseudo-feedback less often.

We used a hierarchical reinforcement learning task adapted from Mas-Herrero et al. (2019) (Figure 3.1). Participants' main goal was to accumulate as many points as possible in a 2-step task. In the first step, participants had to choose, in less than 2 seconds, between two shapes (triangle or square) by pressing either the left (index finger) or right button (middle finger) of a response pad. Then, 1200 ms after the selection, the figure turned out into a colour (pseudo-feedback). Half of the participants were instructed that the blue colour could lead to points (pseudo-reward),

while the green colour was always associated with no points. The other half of the participants received the opposite instructions. Then, 500 ms after the pseudo-feedback was delivered, the selected shape randomly came out on either the left or the right side of the screen. Participants were compelled to indicate the position by pressing the corresponding pad button as rapidly as possible (time limit of 1000 ms). This step was designed to avoid central tendency bias or participants unwilling to answer with extreme responses. Finally, if the shape turned to the positive pseudo-feedback colour (pseudo reward, first step) and the participant correctly responded to the position of the shape (second step), they could be rewarded (25 points) with a certain probability (p2). The feedback image (tick or cross) stayed on the computer screen for 1000 ms, and after a fixation point of 2000 ms, the next trial started (Figure 3.1).

In Experiment 1, the two shapes had the same probability of final reward. One shape (PS+) was associated with higher pseudo reward probability (p_1^{PS+} = 0.6 for group 1; p_1^{PS+} = 0.7 for group 2; p_1^{PS+} = 0.8 for group 3) but lower final reward probability after obtaining a positive pseudo-feedback (p_2^{PS+} = 0.4 for group 1; p_2^{PS+} = 0.3 for group 2; p_2^{PS+} = 0.2 for group 3; Figure 1). In contrast, the other shape (PS-) presented the opposite pattern: less probability of obtaining pseudo-feedback (p_1^{PS-} = 0.4 for group 1; p_1^{PS-} = 0.3 for group 2; p_1^{PS-} = 0.2 for group 3), but higher probability of final reward after obtaining a positive pseudo-reward (p_2^{PS-} = 0.6 for group 1; p_2^{PS-} = 0.7 for group 2; p_2^{PS-} = 0.8 for group 3). It is important to note that the two shapes yielded to the same probability of obtaining a final reward (p_{reward} = .24 for group 1; p_{reward} = .21 for group 2; p_{reward} = .16 for group 3).

The task had 204 trials, among which 68 were free-choice, so participants had to select between one of the two shapes. The rest, 136, were

forced-choice trials, in which only one shape was presented on one of the sides of the screen (forced-choice trials). These forced-choice trials were included to ensure that the two options were sampled equally.

Experiment 2

Participants.

One hundred eighteen healthy university students from the Faculty of Psychology of the University of Barcelona (M = 22.7 years, SD = 6.1, 28 men) participated in the experiment for course points. All participants gave written informed consent, and the Ethical Committee approved all procedures of the University of Barcelona).

Experimental procedure

In Experiment 2, we used the same task structure, number of trials and time framing between the two steps as in experiment 1. However, we wanted to determine to what extent people were able to select the most pseudo-rewarding shape (PS+) at the cost of obtaining fewer final rewards. Group A (N = 56) and group B (N = 62) started with different initial probabilities to get points. Therefore, the probabilities of p1 and p2 were different for each group. Group A started with $p_1^{PS+} = 0.7$, $p_1^{PS+} = 0.3$ / $p_2^{PS-} = 0.3$, $p_2^{PS-} = 0.7$ and group B with $p_1^{PS+} = 0.8$, $p_1^{PS+} = 0.2$ / $p_2^{PS-} = 0.2$, $p_2^{PS-} = 0.8$. The p1 probability of each shape changed after receiving a reward in free-choice trials (1/3 of the total trials). After obtaining a final reward, p1 of the selected shape was reduced by 0.02, and p1 of the non-selected shape was increased by the same amount (Figure 1). All in all, the bias for a shape increases the chances of success by the less selected one. It is important to remark that the p2 did not change along with the task. Therefore, the total

probability of obtaining a final reward (p(reward)) changed with the change of p1 (p(reward) higher in the less-chosen option). This manipulation of the probabilities was not revealed to participants.

Behavioural analysis

Experiment 1

We used a Hierarchical Bayesian Model (HBM) proposed in Krusche (2014) with the STAN R package to determine participants' preferences in experiment 1. HBM models are used to describe the probability of occurrence of a dichotomous event that has multiple dependencies or co-dependencies and links parameters by common distributions. Therefore, the posterior distribution for a parameter is influenced by the results in other parameters and vice versa (Krusche, 2014; Gelman, 2006). In the present experiment, we used three hierarchies of interrelated dependencies. At the bottom, the probability of each participant selecting the less pseudo-rewarding shape in the free-choice trials (θs) was modelled, for each group, using a beta distribution (beta(α_g, β_g)) with a given mode (ω_g) and concentration ($\kappa_g = \alpha_g + \beta_g$)). In addition, the model assumed that all the results of the different experiments came from a global beta distribution (at the top of the hierarchy) with its corresponding mode (ω_0). A prior beta distribution with parameters $\alpha = 4$ and $\beta = 4$ was used, representing an equivalent preference for the two options. κ used a prior gamma distribution that assumes a low uncertainty between the parameters of the model (s= 6.25; r = .125). We used κ to detect extreme preferences among the participants and give consistency to the MCMC during the burn-in period (period of convergence). We ran four MCMC with 2000 iterations, each in 22000 burning steps.

Experiment 2

In the second experiment, the probability of obtaining the final reward changed with the preference for one or another figure. Therefore, we wanted to determine to what extent people assumed a cost of selecting the most pseudo-rewarding option by reducing its probability of final reward. In case of not being biassed towards one of the two options, people should select them in an equivalent way, leading to similar values of p1 at the beginning and at the end of the experiment. In contrast, if people had preferences for one of the two options, this would be reflected in a decrease of p1 for this shape and an increase for the opposite (e.g., a preference for PS+ would decrease p1 for this figure and increase it for PS-), leading to a differential probability of final reward for the two options. In order to study this phenomenon, we computed the point of subjective equivalence (PSE) as the p1 probability for PS- at which the two options were equally selected. If PSE was higher than the initial p1(PS-) (0.3 in group A and 0.2 in group B), this would indicate a preference for the most pseudorewarding shape.

We used a Robust Bayesian Logistic Model as proposed in Krusche et al. (2014) with the JAGS R package to compute this value. The data was composed of the binary response of choosing between PS+ and PS- and the current probability of getting a pseudo reward at selecting PS- as a predictor. Therefore, we used a logistic model with an intercept (β0) using a normal prior (μ = logistic (β ~ dnorm (M=-0.4, S=0.5)), and a slope (β1, weakly informative prior, normal distribution in the form of μ = logistic (β ~ dnorm (M=0, S=1), Krushke, 2014). We also introduced a third parameter ("guessing" coefficient, α) from a beta distribution prior to controlling for the outliers under the assumption that extreme values would be very unlikely (α ~ dbeta (a=1, b=6), Krushke, 2014; Wu and Jermaine, 2007). Monte Carlo

Markovian Chains (MCMC) were run with 20000 iterations, each in 22000 burning steps.

Results

Experiment 1

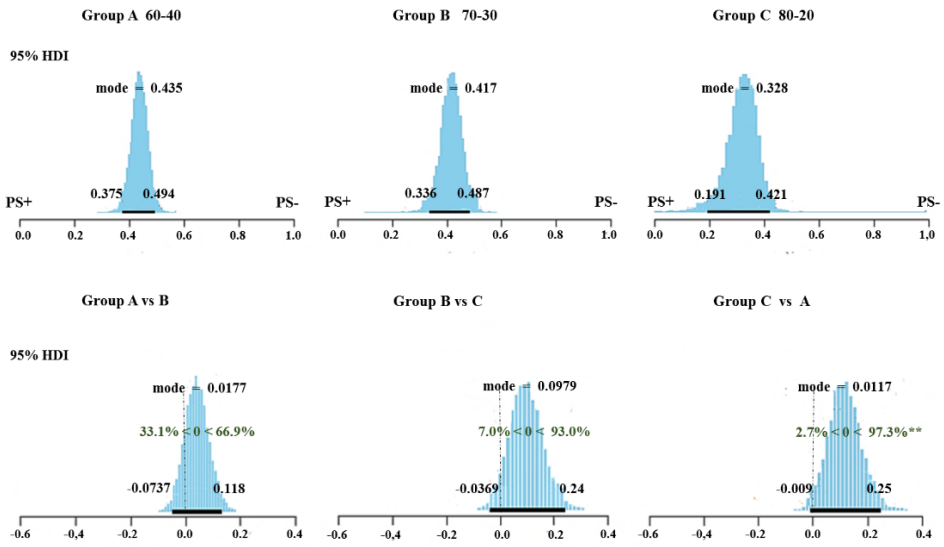


Figure 3.2. Preference across the groups. A. Distribution of posterior probabilities for the preference towards the PS- figure for the three groups. In all the cases, 95% of the highest density intervals are below .5 and did not include 0, indicating that participants consistently selected fewer times the less pseudo-rewarding figure. B. Comparison between the preferences for the PS- figure among the three groups.

The hierarchical model was feasible (see Supplementary Materials in ANNEX) and showed a lower preference for PS- in all groups (Figure 3.2). The bias for the most pseudo-rewarding shape was presented, revealed by a significant probability of selecting PS- less than 0.5 for group 1 (M = .43, HDI = .37:.49), group 2 (M = .41, HDI = .34:.48) and group 3 (M = .33, HDI

= .19:.42). In all groups, 95% of the highest density interval (HDI) was below .5, indicating that participants consistently select the most pseudo-rewarding shape PS+. In other words, participants preferred the option that delivered more pseudo-feedback even when the final reward was the same for the two options.

When the groups were compared, the differences were especially notable in the comparisons between group 1 vs group 3 and group 2 vs group3. Therefore, even when all the HDI contained the 0 value, in the former, 94.6% of the HDI showed a higher preference for PS+ than for PS- in group 3 compared to group 1, suggesting an increased effect in those conditions with higher differences between p1 and p3.

Experiment 2

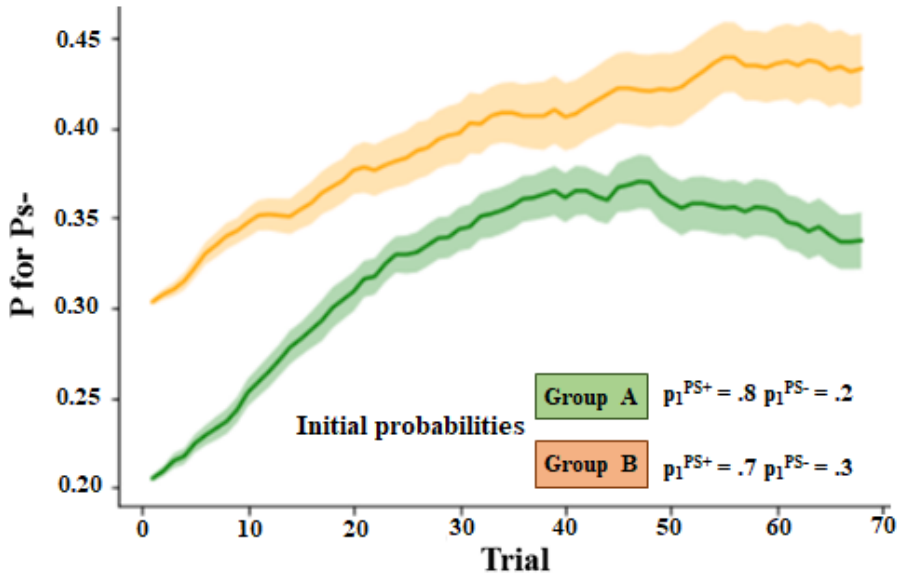


Figure 3.3. Probability of obtaining pseudo-reward for the PS- figure across free trials. PS- refers the option that delivers pseudo-rewards less often. P_1 of the less pseudo-rewarding figure increased 0.02 each time participants received a reward in the PS+ and decreased when they received a reward in PS-. The increase in P1 shown in the two groups indicates that participants consistently selected PS+ throughout the duration of the task.

The results of figure 3.3 show a sustained increase in the probability assigned to the PS- figure (p_1^{PS-}) for groups A and B, indicating that, as in Experiment 1, participants consistently preferred the PS+ option. Thus, at the end of the free trials, the probability of final reward (p_{reward}) for option PS- was higher than PS+. These results suggest that people preferred the option with more pseudo-feedback (PS+), even when it implied a less optimal alternative.

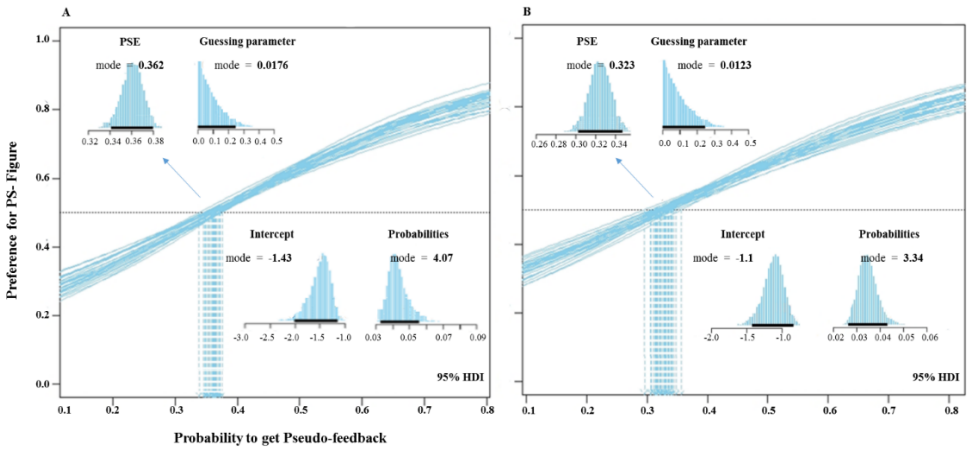


Figure 3.4. Point of subjective equivalence (PSE). Bayesian logistic regression across participants in Group A and B showed the PSE where participants selected indifferently between both options according to p_1^{PS-}. In the two groups, PSE was higher than the initial p_1^{PS-} (where the two options were equally rewarding), showing that participants are willing to assume a cost to select more often the PS+ figure.

The logistic regression for groups A and B allowed the determination of the point of subjective equivalence (PSE) between both options (Figure 3.4). This point (PSE) indicates a moment where there is no biased behaviour toward a particular shape (p = 0.5 for selecting the two figures). In group A PSE was p_1^{PS-} = .36 (HDI = .34:.38) and, in consequence p_1^{PS+} = .64 (HDI = .62:.66). Given that p_2 remained constant throughout all the experiments, the

total probability (p_{reward}) for PS- was 0.252, while for PS+ was 0.192, showing that participants were willing to pay a cost to select the PS+ figure. For Group B, the probability of selecting the less pseudo-rewarding figure was $p_1^{PS-} = .32$ (HDI= .30:.34). This yielded a final probability of reward of $p_{reward} = 0.256$ for PS- and 0.136 for PS+. Again, in this case, participants assumed a cost to bias their decisions towards PS+.

Discussion

In the present study, we aimed to determine if pseudo-rewards could lead to sub-optimal choice behaviours. In a series of experiments, we demonstrated that individuals prefer the most pseudo-rewarding options, even if they are not optimal in terms of ultimate rewards. In simple paradigms, these results demonstrate conclusively that information in intermediate states is reinforcing and may lead to biased behaviour. Intriguingly, a preference for the most pseudo-rewarding figure was found in all of our experiments, both when the two options led to the same final reward (Experiment 1) and when the selection of this figure led to a reduction in the probability of receiving the final reward (Experiment 2).

Experimental settings with secondary goals can be used to observe how the overall complexity of a task is addressed by leading the individual's attention to a smaller set of subroutines that pile up sequentially towards a goal. The decision bias reported in this study outlines the importance of information on the different steps required to complete a task or obtain a final reward, as it has previously been highlighted as a determinant during decision making and learning (Mas-Herrero et al., 2019; Mc Delavitte,

2016). Similar to the intermediate steps considered in the current experiment, these sub-goals could potentially improve the predictive values for the pre-states that anticipate a final reward, therefore speeding up learning and helping to excel goal-oriented behaviours (Diuk et al., 2013a). Importantly, it has been demonstrated that the prediction errors associated with these pseudo-feedbacks engage regions of the reward network such as VS in a manner similar to reward prediction error (Ribas-Fernandes et al., 2011; Diuk et al., 2013b, Mas-Herrero et al., 2019), supporting their reinforcing properties and justifying their engaging nature, which may lead to such sub-optimal behaviours. In this line of research, Mas-Herrero et al. (2019) discovered that the activation of the Ventral Striatum with pseudo-rewards correlated with the preference for PS+ choices, indicating that individual differences in the preference for PS+ may be related to the differential impact of pseudo-rewards in the reward neural circuitry.

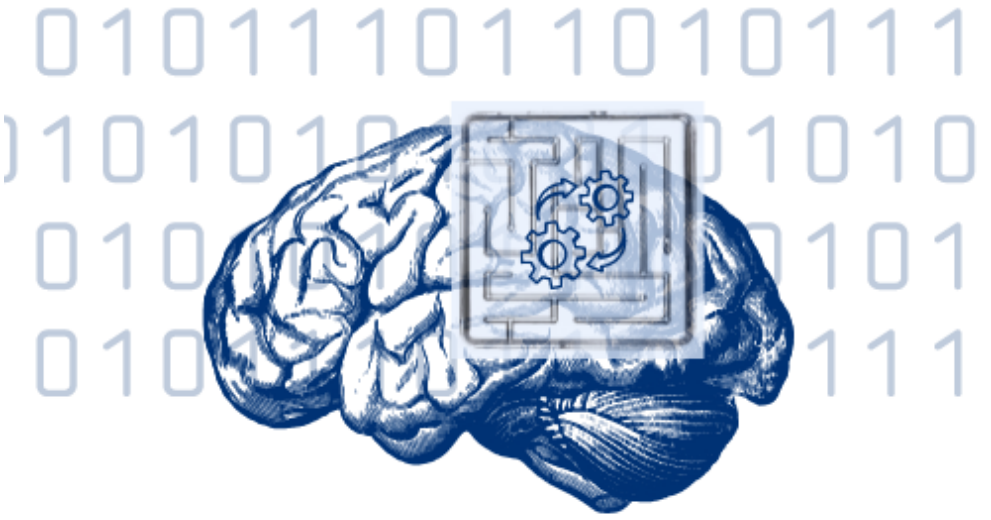
Different explanations assert that people are directed by statistically optimum decision-making (Bogacz, 2007), but the results of the current study imply that information on intermediate steps may bias decisions toward poorer decisions. These findings are coherent with prior research that has demonstrated sub-optimal choice behaviours and maladaptive decision-making in pigeons (Stagner, and Zentall, 2010; Spetch et al., 1990), rats (Chow et al., 2017; Lesaint et al., 2015), and humans (Colaizzi et al., 2019). In all these cases, the subjects were studied while making decisions involving more salient stimuli or decisions with higher predictive value at early states in the task, which presumably, could lead to obtaining lower rewards. Therefore, the sub-optimal behaviours demonstrated by the current experiment's participants suggest that pseudo-feedbacks operate as discriminative stimuli that overshadow other possibilities, making the information of Ps+ more prominent and omitting the information coming

after picking PS-. Considering pseudo-feedbacks as necessary informative cues, present results suggest that sub-goals might acquire a leading role in inducing decisions throughout the task. In this sense, pseudo-feedbacks would be desirable in a similar way as the conditioned stimulus in sign-tracking procedures, not only because they are needed for the accomplishment of the final goal, but because they become attractive and elicit approach behaviours towards them. In this sense, pseudo-feedbacks would be valued in the same manner as the conditioned stimulus in sign-tracking procedures, not only because they are necessary for the achievement of the ultimate goal but also because they become attractive and motivate approach behaviours.

Indeed, different animal (Bromberg-Martin and Hikosaka, 2009, Anderson et al., 2015) and human (Eliaz and Schotter, 2010) studies have proposed that information is rewarding per se and that some biases or even sub-optimal behaviours might be explained by the preference for informative over non-informative stimuli. While it could be argued that participants preferred the PS+ stimulus due to its informative value, both PS+ and PS- provided the same amount of information. In addition, in most situations, PS- resolved earlier the uncertainty, discarding the role of uncertainty avoidance (Kreps and Porteus, 1978; Sharot and Sunstein, 2020) in the preference for PS+. In addition, the fact that the time needed to obtain the final reward was the same in the two options (even in the case of a negative pseudo-feedback) also discarded the possibility of a preference for one option due to temporal delay discount (Kirby et al. 1999; Rung and Madden). Finally, participants' preferences cannot be related to imbalances in the presentation of the stimuli (Niv et al., 2002). Therefore, even when participants had a clear preference for the most pseudo-rewarding figure, as shown in the two experiments, they had access to the contingencies of the

two options in the forced-choice trials (1/3 of trials for each shape). Therefore, participants were exposed to the contingencies of the two keys a similar number of times, avoiding the possibility of biases due to different expositions to the stimuli.

In conclusion, the current study demonstrates that pseudo-rewards have a reinforcing nature, which can lead to sub-optimal decision-making during the experimental task. Future research examining the limits of this bias utilising more extreme probability differences is required to comprehend the capacity of pseudo-reward to influence decisions.



Chapter 4. Study 2

Chapter 4. Study 2. Theta Oscillatory activity during sub-optimal choice behaviours

Summary

Individuals divide complex behavioural routines into a series of steps to facilitate learning. Each step might provide information about their accomplishment. These pseudo-rewards might bias the decisions towards them, leading to maladaptive behaviour. However, the oscillatory components underlying these sub-optimal decisions are poorly understood. To study this phenomenon, EEG was recorded from twenty-four healthy subjects who participated in a two-step task. In this experimental paradigm, participants could select between two options, one providing more pseudo-rewards than the other one but with the same probability of obtaining the final reward. However, as the task advanced, the probability of obtaining the final reward from the most selected option decreased. Three different reinforcement learning models were used to model the behavioural data, and their results were used to study the oscillatory activity associated with reward and pseudo-reward prediction errors. Results showed that participants consistently selected more the most pseudo-rewarding option, even when this yielded to a sub-optimal behaviour in the form of less final rewards. In addition, theta oscillatory activity at frontal electrodes was associated with both reward and pseudo-reward prediction errors. This data provide evidence for the critical role of theta activity in sub-optimal behaviours.

Introduction

Decision-making and learning interact across time and are essential in human behaviour. In day-to-day life, individuals decide among a set of options based on their situation and previous learning. The outcomes of such decisions impact future choices, updating the value of each option and allowing a flexible adaptation of behaviour (O'doherty et al., 2017; Glimcher and Fehr, 2013). Under the Reinforcement Learning framework (RL; Sutton and Barto, 1998), this update is performed using the prediction error (PE) of the performed actions, that is, the difference between the expected and obtained results in a specific time frame (Diuk et al., 2013a). Therefore, successful selection action to get positive outcomes would reinforce the same decision, and negative outcomes would encourage the search for alternative choices. Therefore, the standard RL framework and its correspondent behavioural models have been demonstrated to be a reliable approach to explaining human behaviour in simple settings. However, complex (and more realistic) scenarios involve sequences of actions necessary to get the final reward and are challenging for standard RL algorithms (Botvinick et al., 2009). Such a series of actions are taken to reach different steps which, when carried out correctly, lead to a final reward. Standard RL considers a single level of decisions, with a direct link between action and reward, but might fail to appropriately model these situations where multiple choices might interact for an expected outcome (Botvinick and Weinstein, 2014; Threadgill and Gable, 2018), yielding to a limitation of these approaches to explain more naturalistic phenomena (Botvinick et al., 2015).

However in situations in which long chains of decisions are needed to reach a final goal, critical information which speeds up learning is the

knowledge of the correctness of the different actions taken in each step or sub-goal (O'Doherty et al., 2015). Therefore, in order to reduce uncertainty, individuals break down the tasks into progressive steps or states until the final goal attainment. This implies that as the task unfolds, individuals must reach a series of sub-goals hierarchically organised (Balaguer et al., 2016; Ribas-Fernandes, Niv and Botvinick, 2011), receiving in every one of them, information (pseudo-feedback) about whether they have accomplished this sub-task. The process of learning by action selection stages has been coined hierarchical reinforcement (HRL), and it proposes a solution to the temporal problem of strictly fixed experimental conditions. HRL works by escalating and grouping by levels of the different options and steps, assigning PEs and predictive values to each condition at each state (Botvinick et al., 2012; Holroyd and Yeung, 2012; Konidaris and Barreto, 2009).

The study of the neural correlates of the prediction errors generated in the different sub-steps of the task (so-called Pseudo Reward Prediction Errors, PRPE) has revealed the involvement of similar brain regions to those computing reward prediction errors (Ribas-Fernandes et al., 2011; 2018; Mas-Herrero et al., 2017), including anterior cingulate cortex (ACC; Ribas-Fernandes et al., 2011; Alexander and Brown, 2019), ventral striatum (VS; Balleine and Delgado, 2007; Schönberg et al., 2007; Garrison et al., 2013) and medial prefrontal cortex (mPFC; Chiang and Wallis, 2018; Collins, 2018; Holroyd and Yeung, 2012; Ribas-Fernandes et al., 2019). In addition, previous studies have also described an important role of theta oscillatory activity in the computation of reward prediction errors (Cavanagh et al. 2010, Mas-Herrero et al. 2016), advocating the notion that this activity would be a predictor of learning. However, little is known about the oscillatory neural activity of PRPE and whether theta oscillatory activity is the main neurophysiological mechanism involved in its processing.

The increased number of studies on the computational properties of HRL and the associated neural correlates of PRPE and RPE notwithstanding, less attention has been devoted to one of the most striking properties of pseudo-rewards, the capacity to bias agents' behaviour by focussing on pseudo-rewards rather than rewards. Indeed, given that accomplishing sub-goals is a prerequisite for a final reward, obtaining pseudo-rewards might end up being a goal by itself. Such a phenomenon might be possible if every time there is an adjustment to PRPE, there is a reinforcement of a stable preference for alternatives that facilitate the attainment of sub-goals. In a recent study, Mas-Herrero et al., 2019 showed that, when faced with two decisions with the same probability of final rewards, people prefer the one with more pseudo-rewards. Therefore, the engaging properties of pseudo-rewards could bias the decision, even in conditions when they would lead to the worst outcome or in situations where sub-goals are more difficult to acquire but closer to the final rewards (Collins, 2018; Ribas-Fernandes et al., 2019: Mas- Herrero, et al., 2019). The engagement with positive pseudo-feedbacks might stop the exploration of new options in a sort of attentional bottleneck and information avoidance (McGovern and Barto, 2001; Pateria et al., 2021). Thereby, a persistent bias for alternatives that deliver more pseudo-feedback might mean that individuals could assume a cost to the point of lessening their chances of obtaining a final reward.

The goal of the present research is to study the oscillatory mechanisms underlying the processing of pseudo-reward prediction errors. We used a modification of the experimental paradigm of Mas-Herrero et al. 2019, in which participants face a two-step task with different probabilities of obtaining a pseudo-reward. Importantly, as the task unfolds, the final reward of the most selected option will decrease, allowing the study of the neural mechanisms underlying the bias towards the most pseudo-rewarding

option. We hypothesise that theta oscillatory activity will play a key role in the processing of both reward prediction errors and pseudo-reward prediction errors.

Methods

Participants.

Twenty-four healthy university students from the Faculty of Psychology of the University of Barcelona (M = 22.7 years, SD = 3.1, 9 men) participated in the experiment for 20 € an extra monetary reward based on their performance in the task. Every successful trial added 0.20 € real money to the final amount. All participants gave written informed consent, and the Ethical Committee of the University of Barcelona approved all procedures.

Task

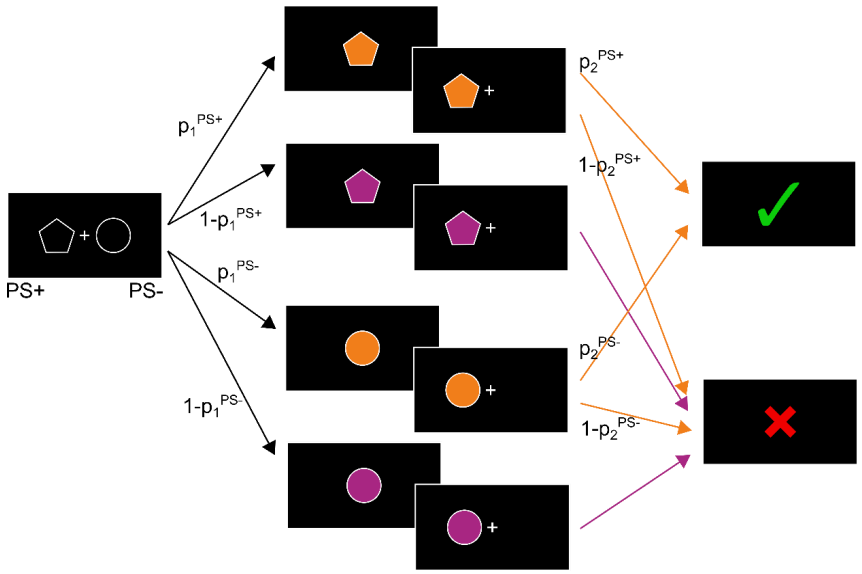


Figure 4.1. Decision-making task used. The task consisted of two steps. First, a certain colour (orange in the present example) had to be obtained to have the opportunity to gain money in the next step (green tick). In free-choice trials, participants had to choose between two options, one figure (PS+), which had a higher probability of obtaining pseudo-rewards

(p_1^{PS+}) but a lower probability of obtaining final rewards after positive pseudo-feedback (p_2^{PS+}). The other figure (PS-) had the opposite pattern. Initially, the two figures had the same probability of obtaining a final reward, but p_1 changed along with the task, leading to differences in the final reward probability (p_1 x p_2). In free-choice trials (1/3 of the choices), participants could select between the two figures. In the rest of the trials (2/3, forced-choice), only one figure was presented on the screen.

A decision task adapted from Mas Herrero (2019) was used (Figure 4.1). This experimental paradigm contains two hierarchical levels, where a first sub-goal needs to be accomplished (pseudo-feedback) to reach the final reward (feedback). The participants were instructed to try to get as many final rewards as possible to increase their final pot. At the beginning of each trial, the participants chose between two geometric shapes within two seconds (Shape 1 or Shape 2). The choice was made by pressing the right or left button on a pad. Each figure was assigned an initial probability p_1 to provide a positive (PS +) or negative (PS-) pseudo-feedback represented by the chosen figure painted in a particular colour. If participants received a certain colour (blue or orange in half of the participants, green or purple in the other half), they had the chance to win money at the end of the trial (pseudo-reward). If they received the other colour (green or purple, or blue or orange figure respectively), they would lose the trial for sure.

After this, the shape appeared on the left or the right side of the screen, and the participant had to select the correct side by pressing the corresponding button. In all the cases in the whole Block A, there was a colour associated with the probability of obtaining points and another to get no points. The combination of colours randomly changed at the beginning of Block B. Finally, if the participants had received positive pseudo-feedback in the first step, they had a certain probability (p_2) of obtaining a reward (0.2 €, green tick) or no reward (1- p_2, red cross). In contrast, in the case of receiving a negative pseudo-feedback in the first step, the final feedback was

always no reward (red cross). The feedback remained on the screen for 1000 ms, and then a fixation point was presented for 2000 ms before the next trial. The experiment was divided into two blocks (A and B) in which the participants performed the same task twice, but with different probabilities, shapes and colours. Figures and colours were counterbalanced within participants and blocks.

At the beginning of each block, the two shapes had the same chance of winning at the end of each trial (p_1 x p_2). Shape 1 (PS +) was associated with a higher probability of receiving colour PS+ (p_1(PS+) = 0.7 for block A; p_1(PS+) = 0.8 for block B), but a lower probability of rewarding feedback after receiving a positive pseudo-feedback (p_2(PS+) = 0.3 for block A; p_2(PS+) = 0.2 for block B; Figure 4.1). In contrast, Shape 2 (PS-) presented a lower probability of receiving positive pseudo-feedback (p_1(PS-) = 0.3 for block A; p_1(PS-) = 0.2 for block B), but higher probability of rewarding feedback after receiving a positive pseudo-feedback (p_2(PS-) = 0.7 for block A; p_2(PS-) = 0.8 for block B; Figure 4.1). Therefore, in the two blocks, participants began the task with the same probability of receiving a final reward regardless of their preference for a shape (p(reward) = .21 for Block A; p(reward) = .16 for Block B). However, the probability p_1 of each shape changed after receiving a final reward in the Free Choice trials, decreasing 0.02 for the chosen form and increasing 0.02 for the non-chosen form. This manipulation allowed us to study to what extent participants would perseverate in one of the options, even when the final probability of reward (p_1 x p_2) was reduced for this shape and increased for the alternative one. The participants were not informed about this manipulation.

The task consisted of 204 trials in each block. In 68 trials, participants had to select one of the two shapes (free choice trials). In the remaining 136 trials

(forced trials), only one shape was presented on one side of the screen, and the participant had to select it by pressing the corresponding button (68 trials for each shape). The order of the forced and free trials was randomised. Thus, the participants had the chance to learn the probabilities of the two shapes independently from their decisions in the free-choice trials.

Temporal difference Models

We tested three different temporal difference learning models (TD0) in order to find the one that better suits the behavioural evidence. First, we tested a model with fixed values for learning rate (no free parameters, $\alpha = 0.5$ for pseudo-feedbacks and feedbacks, model 1). Second, we used a model with two learning rate parameters, one for pseudo-feedbacks α_1 and another one for final feedback (α_2, model 2). Finally, in the third model, we considered three different free parameters, two learning rates and a parameter that indexed the value of pseudo-feedbacks (model 3). All models included an extra parameter beta, which was the one to adjust the model due to a softmax function (see below).

The action values of the model were based on the history of pseudo-feedbacks (V_1) and feedbacks (V_2) and updated according to the reward prediction errors. Two RPEs were computed in this experiment: when the pseudo-feedback was delivered (RPE1) and when the final outcome was achieved (RPE2). RPE1 was modulated by the pseudo-feedback, indicating whether the sub-goal had been accomplished or not. In parallel, the final reward and $V_{1,t}$ determine RPE2, and $V_{2,t}$ along with the trials. All the parameters and variables were calculated per subject.

RPE1. The action value $V_{1,t}$ at trial t of each shape (PS+; PS-) was updated when the pseudo-feedback was presented:

$$V_{1,t+1} = V_{1,t} + \alpha_1 * RPE1_t \quad (3.1)$$

being α_1, the learning rate associated with the Pseudo-feedback. As stated above, in the case of Model 1, the learning rate was the same for RPE1 and RPE2 fixed at $\alpha_1 = 0.5$, while for the Models 2 and 3, α_1 was a free parameter (for details on parameter optimisation, see below). RPE1 was computed as:

$$a)\ RPE1_t = V_{2,t} - V_{1,t} \quad \rightarrow Model\ 1\ and\ 2 \quad (3.2)$$

$$b)\ RPE1_t = V_{2,t} + pr - V_{1,t} \quad \rightarrow Model\ 3$$

in case of positive pseudo-feedbacks, and $RPE1_t = -V_{1,t}$ for all the models in case of negative pseudo-feedback. The parameter *pr* included in model 3 represented a fixed value of receiving positive pseudo-feedback and was introduced as a free parameter and fitted individually for each participant. The variable $V_{i,t}$ represents the action value of each figure, and the subscript *t* represents the trial number.

RPE2. The action value $V_{2,t}$ of each shape delivering positive pseudo-feedback was updated when the feedback was presented, while the value for negative pseudo-feedback remained constant:

$$V_{2,t+1} = V_{2,t} + \alpha_2 * RPE2_t \quad (3.3)$$

The action values of the two figures had the same initial value for the two shapes. In the case of Model 1, α_2 was fixed as a learning rate = 0.5, while for the model, the two learning rates were the same ($\alpha_2 = \alpha_1$) and were free parameters. In the case of model 3, α_1 and α_2 were independent free parameters. RPE2 was computed as:

$$RPE2_t = r - V_{2,t} \quad \rightarrow Model\ 1,\ 2\ and\ 3 \quad (3.4)$$

Where *r* was +1 or 0 depending on whether the subject received money or not.

Model Fitting

In order to fit each model to subject behaviour, we assumed that subjects made their choices stochastically according to a soft-max function, dependent on the experienced value.

$$P\left(\hat{resp}_t = j\right) = exp(\beta V^{j}_{1,t}) / \sum_{i} exp(\beta V^{i}_{1,t}) \quad (3.5)$$

where j corresponds to the different shapes. The value of the parameter β ® specifies the steepness of the decision function. The value of β was assigned individually to subjects and was, along with other model parameters, found by maximising the log-likelihood of the model performing the same response as the subject across all *n* trials

$$\hat{\theta} = argmax \sum_{t=1}^{n} log\ P(\hat{resp}_t = resp_t) \quad (3.6)$$

where the set of individually optimised parameters was θ={} for model 1, θ={ α_1, α_2 } for model 2 and θ={α_1, α_2 and pr} for model 3. In addition, the β value was fitted. The optimisation was done using Matlab's fminsearch function (Mathworks, Massachusetts).

EEG recording and analyses

Electroencephalogram (EEG) was recorded using a BrainAmp amplifier (Brain Products GmbH; band-pass filter: 0.01–100 Hz, with a notch filter at 50 Hz and 250 Hz sampling rate) with tin electrodes mounted in an elastic cap with 29 electrode standard positions (Fp1/2, Fz, F7/8, F3/4, FCz, FC1/2, Fc5/6, Cz, C3/4, T7/8, Cp1/2, Cp5/6, Pz, P3/4, P7/8, Po1/2, Oz). Throughout

the experiment, electrode impedances remained below 5 kOhms. Four external electrodes were employed, including one electrode placed at the lateral outer canthus of the right eye used as an online reference, one electrode placed at the infraorbital ridge of the right eye to control eye movements, and two electrodes in left and right mastoids. Participants were instructed to not blink while receiving either feedback or pseudo-feedback and to try to blink during the fixation periods. Epochs with the different pseudo-feedback and feedback conditions were extracted from −2000 ms before the feedback stimuli to 2000 ms after it.

Trials with amplitudes higher than ±100 μV between –100 ms and 1000 ms were excluded. Three subjects presented more than 30% of the trial rejection rate across the whole task and were excluded from the analyses. In order to analyse the time-frequency of the signals, trials were convolved with complex Morlet wavelet analyses for the different epochs, scaling them from 1 to 40 Hz. Changes in time-varying power respect baseline (100 ms before stimuli) were computed for each trial and averaged for each subject; then, a grand average was performed.

Statistical Analysis.

We analysed the power changes at the Fz electrode, which has been shown to be maximal in the frontocentral theta activity associated with positive and negative feedback, as well in RPE computation (Mas-Herrero and Marco-Pallarés, 2014). Repeated-measures ANOVA was computed with three within factors: valence (positive or negative feedback or pseudo-feedback), shape (PS+ or PS- figure) and block (A or B). In addition,

in order to study the relationship between theta power and RPE or PRPE, we computed two Linear Mixed Effect Model (LMM) using as independent variables the Reward Prediction Errors RPE1 and RPE2 produced by the temporal difference model introduced to explain learning and decision along with the task. We assumed different random intercepts for each subject and trial and considered both trial and subject-level variance in the participants' performance. The model selection was based on the Akaike criterion (AIC) using the lmer R package (Bates et al., 2017)

Results

Behavioural results

Participants presented on average a preference for the PS+ figure during the Block A (M = 0.65, STD = 0.14), which was significantly above the randomness scenario of 0.50 ($t(23) = 5.0$, $p < 0.001$)), and was replicated in the Block B (M = 0.64, STD = 0.13), which was significantly above the randomness scenario of 0.50 ($t(23) = 5.3$, $p < 0.001$)). This result shows how the bias towards the figure which delivered pseudo rewards with higher frequency was present across trials and the blocks. On average, the number of rewarded trials was similar in Block A (M = 28.6, STD = 4.6) and Block B (M = 28.4, STD =3.9), which meant a monetary compensation of €5.73 and €5.68, respectively.

These final results suggest overall probabilities of getting a reward ((p(reward)) of 0.14 and 0.13 for Block A and B, respectively. However,

compared with an optimal scenario for the task where participants would hold constant odds to obtain final rewards (p(reward) = .21 for Block A; p(reward) = .16 for Block B), the behavioural data showed a significant sub-optimal behaviour. Indeed, block A was significantly lower than the optimal scenario ($t(23) = -15.1$, $p < 0.001$), as well as for block B ($t(23) = -5.3$, $p < 0.001$). The behavioural data indicates how the preference for PS+ was prominent even at the cost of earning less rewarded trials, therefore less money.

Model Comparison

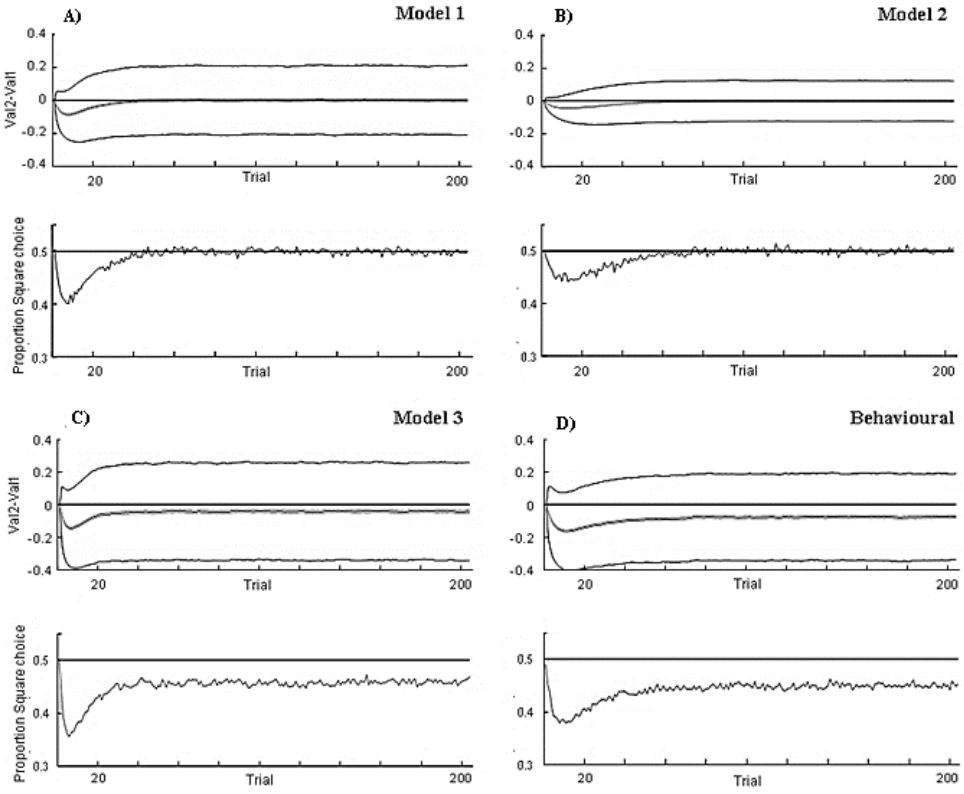


Figure 4.2. Reinforcement Learning Models. Differences between the V1 value between PS- and PS+ and the difference between probability choices provided by model 1 (A), model 2 (B) and model 3 (C). Parameters for the models were fitted to individual subject data, and an average parameter set was used for the plot. Simulations of choices are done

10000 times for each model (204 trials per simulation), and the plots were averages (+- standard error). 2D) Simulation-based on model 3 (C above), with parameters fitted to individual choices based on real subject behaviour.

In order to compare the three models, the parameters for each model were fitted to individual subject data, and an average parameter set was used (Figure 4.2). Simulations of choices were done 10000 times for each model with 204 trials per simulation as in real experiments. As can be seen in Figure 4.2, the inclusion of two learning rates (α_1 and α_2) in model 3, in contrast to a fixed value for model 1 (Figure 4.2A) and the same values for two learning rates in model 2 (Figure 4.2B), slowed down the returning to the equal selection to the two options. In addition, the inclusion of a value for the pseudo-reward (pr in model 3) maintained a preference for PS+ along with all the tasks (Figure 4.2C), which was similar to the one observed in the participants (Figure 4.2D). The parameters that best fitted the behavioural results for model 3 were Pr = 0.47 ± 0.07, α_2 = 0.43 ± 0.19 ; α_1 = 0.39 ± 0.09.

Therefore, this model better reflected the behavioural evidence regarding the preference and bias for the shape that delivers positive pseudo-feedback more often. The Figure 2D shows the parameters of model 3 based on real subject behaviour. In addition, we compared the three models (A, B and C) using a Bayesian model comparison (Penny et al., 2010) to test the models' fit. This analysis outputs an exceedance probability to the event that one model accounts better for behavioural responses than the others. Model 3 received the highest exceedance probability (xp = 0.91), thus being the most likely to describe the log evidence in a subject-by-subject approach (M = 32.78; SD = 6.1).

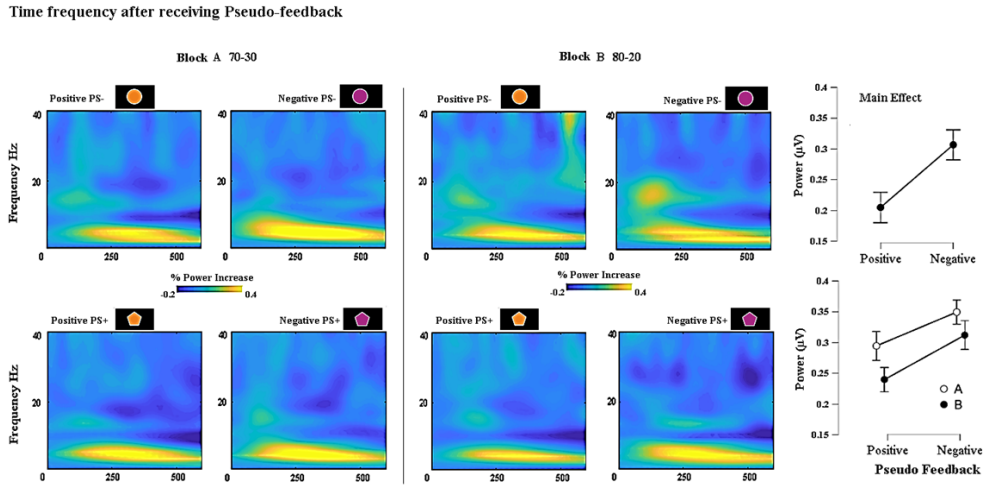


Figure 4.3. Power Increased after pseudo-feedback (0-600ms). The time frequencies were presented per valence (Positive or negative pseudo-feedback) and figure (PS+ or most pseudo-rewarding, PS- least pseudo-rewarding). The left panel indicates the oscillatory activity for the Block A, and the right panel the Block B. The upper figures highlight a significant theta oscillatory activity between 4 Hz to 8 Hz. Main effects are shown on the right side. The top right side shows the main effect of valence in the 4-8 Hz, 100-400 ms time-frequency range. The bottom right side shows the main effects of valence in the two different blocks in the same time-frequency range.

Figure 4.3 shows the time frequency for the different pseudo-feedback conditions at the Fz electrode. All of them showed a clear theta increase (4-8 Hz) between 100 and 400 ms. Repeated measures ANOVA with valence (positive vs negative pseudo-feedback) figure (PS+ vs PS-) and the block (A vs B) revealed only a main effect of valence ($F(1,20) = 12.64$, $p = 0.002$), being the negative pseudo-feedbacks larger than the positive ones (Figure 3B).

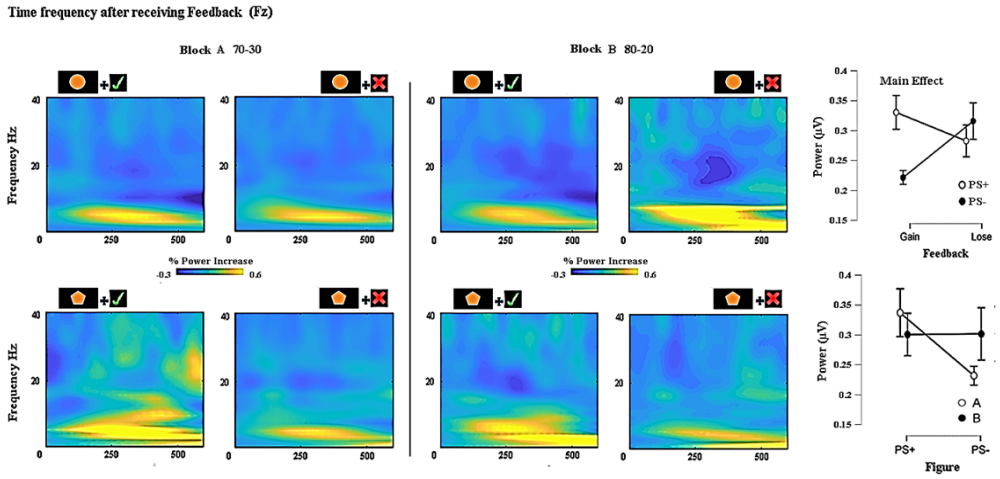


Figure 4.4. Power Increased after feedback (0-600ms). Time frequencies for gains and losses after positive pseudo-feedback per figure (PS+ most pseudo-rewarding, PS- less pseudo-rewarding) and block. The main effects are shown on the right side. The top right side graph shows the main effect of the interaction between figure and valence in the 4-8 Hz 100-400 ms time-frequency range. The bottom right side graph shows the main effect of the interaction between figure and block in the same time-frequency range.

In addition, Figure 4.4 shows the activity after positive and negative final feedback in those conditions in which participants had previously received a positive pseudo-feedback. Similarly to previous results, all conditions showed an increase in theta power in a similar time range to the pseudo-feedback. Repeated-measures ANOVA using the same three factors revealed a significant valence x figure interaction ($F(1,20) = 14.1$, $p = 0.0012$). Power values of Figure 4B show a reversed pattern in the two shapes, with an increase of power for gains compared to losses in the most pseudo-rewarding figures and a decrease in the less pseudo-rewarding one. In addition, rm-ANOVA also revealed a significant and figure x block interaction ($F(1,20) = 5.95$, $p = 0.024$). The figure on the right side bottom shows that, while theta power decreased in the block A (70-30 condition), it did not for the block B (80-20 condition)

In order to study the relationship between theta power and prediction errors (calculated from model 3), two different LMMs were computed for the pseudo-feedback and feedback conditions. Results showed a significant relationship of theta power with the pseudo-reward prediction error ($t(7704) = 2.47$, $p = 0.013$) and reward prediction errors ($t(5698) = 6.34$, $p < 0.001$) computed using RL model 3, supporting the role of theta activity in the computation of the discrepancy between expected and real outcomes both in the intermediate and final steps of the task.

Finally, we also explored the possibility of the existence of a relationship between theta activity and individual differences in the parameters of the model. To this end, we conducted correlations between the theta power of the different conditions and the α_1, α_2 and pr parameters. No significant differences were found for any of these correlations.

Discussion

The present experiment aimed to observe the effect of pseudo-feedback in biased decision-making and its corresponding brain oscillatory activity through an experimental paradigm based on an HRL approach. To this end, a temporal difference model was proposed to explain the behavioural tendencies of the participants and their possible relationship with the oscillatory theta activity. The present results show the involvement of theta oscillatory activity both in the processing of pseudo-feedbacks and feedbacks, as well as for the computation of pseudo-reward and reward prediction errors, reinforcing the key role of this oscillatory activity in HRL.

The first main result of the present study was the preference for the option which provided pseudo-feedback more often (PS+). Although this was previously shown in the original paradigm (Mas-Herrero et al., 2019), in the present study, the repeated selection of one option led to a reduced final reward. These findings are consistent with the proposition that pseudo-rewards modulate learning and, consequently, participants' choices (Ribas-Fernandes et al., 2019; Mas-Herrero et al., 2019). In fact, the rewarding properties of positive pseudo-feedbacks could be explained by an observation effect (Beierholm et al., 2010) and, therefore, PS+ option would generate a motivational effect towards its selection, prevailing even to the larger probability associated with the PS- final reward. Hence, beyond the possibilities of the sole policies' impact in benefiting PRPE accuracy and temporal abstractions to enhance high-level learning (Konidaris and Barreto, 2009; Collins and Cockburn, 2020), the findings in this research suggest a whole behavioural trajectory directed by pseudo-feedback that are not attributed to conventional temporal difference processing.

Interestingly, among the three models tested, the one that best fitted the behavioural data was model 3, which included two different learning rates (one to update the values at the pseudo-feedback level (α_1) and the other for the rewards, α_2), and a third parameter which gave a value to the pseudo-reward. This value was 0.47, indicating that the positive pseudo-feedback receives a specific weight different from the final reward. In addition, such preference for PS+ supports the idea that the more the value given to the positive pseudo-feedback, the more the selection to the most pseudo-rewarding option, even at the cost of obtaining a less final reward. Therefore, as hypothesise, the observed sub-optimal behaviour would be explained by the fact that pseudo-rewards also have a particular value (given,

in the case of our model, by the pr parameter) and, therefore, participants' behaviour is not only guided by actual rewards.

In addition, time-frequency results showed that theta oscillatory activity was critically involved in the feedback processing in the present study. This activity has been proposed to be generated in the Anterior Cingulate Cortex (ACC) and is related to the encoding of reward prediction errors (Holroyd and Yeung, 2012; Gruber et al., 2013; Mas-Herrero and Pallarés, 2016; Shahnazian et al., 2018). However, to our knowledge, no previous studies have been devoted to studying oscillatory activity associated with pseudo-reward prediction errors. Indeed, in our study, theta activity was larger for negative than for positive pseudo-feedbacks (Figure 4.3) and was related to RPE1. Given that PS+ was associated with higher probabilities of receiving positive pseudo-feedbacks (and PS- with less changes of receiving it), it is worth considering the reason why the two figures presented similar behaviour in theta activity. Therefore, at first glance, one might consider that a positive pseudo-feedback in PS- should always be associated with a positive prediction error, as most pseudo-feedbacks are negatives. However, in contrast to RPE2, which only takes into account the final outcome and the V2 value, RPE1 not only takes into account the pseudo-feedback but also the action value of the final feedback. Therefore, being most of the final outcomes negative, the prediction error in the pseudo-feedback dynamically changes on the bases of the history of rewards and pseudo-rewards. Importantly, as stated above, theta power was also found to be associated with the pseudo-reward prediction error. Therefore, the increased activity for the negative compared to the positive pseudo-feedback is compatible with the role of theta in prediction error computation (Cavanagh et al., 2011,

Mas-Herrero et al., 2014) and, in addition, fits well with previous literature showing theta increase with negative feedbacks and loss-related learning (see Glazer et al. 2018 for a review).

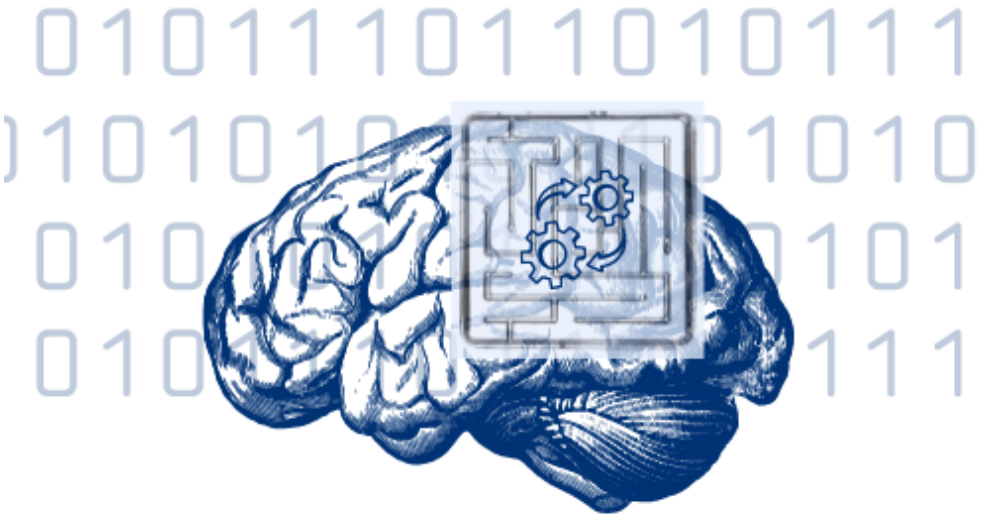
Results on the final feedback, in contrast, revealed an interaction between outcome (gain vs lose) and figure, with increased power in gains vs loss in PS+ and the opposite pattern in PS-. Again, this result goes in line with the role of theta activity indexing reward prediction error. Therefore, given that the probability of obtaining final reward (after a positive pseudo-reward) is much lower in PS+ than in PS-, gains are unexpected in PS+, and losses are unexpected in PS-. Theta activity follows the same pattern and, in addition, correlates with prediction error in the two figures. Overall, both in the pseudo-feedbacks and feedback, results show a clear relationship between theta activity and prediction error, as previously reported using other paradigms (see, for example, Cavanagh et al. 2011, Mas-Herrero et al. 2014).

Importantly, these results also support the critical role of ACC in the temporal abstraction of HRL (Balaguer et al., 2016; Holroyd et al., 2018; Ribas-Fernandes et al., 2019), although its exact role is still under debate. Several accounts have been proposed to explain the role of this area (and its associated neural correlates, in particular, theta activity) in performance monitoring. Our data could fit well with those accounts relating this activity to a surprise of an event (PRO model, Alexander and Brown, 2011) or those suggesting the involvement of ACC in reward prediction error (Holroyd et al., 2016), but not by theories proposing a role of this area only in errors or negative (worse than expected) prediction errors (Holroyd and Coles, 2002). However, it is not clear whether this oscillatory activity plays a role in the bias towards the most pseudo-rewarding option, as similar activations have been found in other studies with unbiased or optimal choices (Cavanagh et

al., 2010; Gruber et al., 2014; Cohen, 2017). Indeed, we found no significant differences between the theta power and the parameters of the model. This might suggest that this mechanism is related to the performance monitoring mechanisms but not to those implied in the bias towards pseudo-rewarding options. In contrast, in an fMRI study, Mas-Herrero et al. 2019 showed that the activation of the ventral striatum with pseudo-reward prediction errors was associated with the preference for PS+ in a non-biased scenario.

In conclusion, these results suggest that, even when these different structures (ventral Striatum and ACC) compute pseudo-feedback prediction errors, they might play different roles in HRL. Indeed, in the classical actor/critic architecture (Takahashi et al., 2008; Hayden and Niv, 2021), the ventral striatum would play the role of the *critic,* evaluating the goodness of the selection in the long run on the bases. This information would be used by the *actor* (dorsal striatum and dorsolateral prefrontal cortex) to select the appropriate action. However, the role of ACC in this model is less clear/straightforward. Some accounts have proposed that, under this architecture, ACC would be related to the selection among different options and the maintenance of the task. Therefore, ACC would be responsible for the learning of high-level options (using and integrating the information provided by the *critic*, probably via VTA/SN) in contrast to the striatum, which would be more oriented to respond to the value of the different events (Holroyd and Yeung, 2012; Holroyd and Umemoto, 2016). Therefore, under this premise, it would be reasonable to assume that, while both VS and ACC (indexed by theta power) activities reflected both pseudo-reward and reward prediction errors, only the *critic* would be related to the value given to the pseudo-rewards. All in all, the present results support the capacity of pseudo-rewards to bias behaviour towards sub-optimal choices and the

critical role of theta power activity in the computation of both pseudo-reward and reward prediction error.



Chapter 5. Study 3

Chapter 5: Study 3. Striatal Contribution during Sub-optimal Decision Making: A hierarchical reinforcement learning approach

Summary

Rewards can be defined as desired results that influence and encourage the behaviour. It is critical to understand how our brain encodes our reward history, resulting in the formation of habits that influence goal-directed behaviour. The more steps required to obtain the reward, the more difficult it is to explain the behaviour. Algorithms based on Hierarchical Reinforcement Learning (HRL) suggest progressive learning in phases, breaking down decision chains into individual actions. This method simplifies the behaviours by stating that completing each step results in the attainment of a final reward. However, completing these stages would represent an objective itself, generating expectations about the outcome, and no previous studies have been conducted on the neural networks involved in the ability of sub-goals to direct learning toward sub-optimal decision-making. Research has suggested that activation of the striatum is critical for concurrent processing of both reward prediction errors and pseudo-reward prediction errors. The purpose of this study was to examine the ventral striatal representations (VS) that underpin pseudo-feedback processing during decision-making in sub-optimal settings. Nineteen university students participated in an fMRI study in a two-step task in which participants could choose between two options, one delivering more pseudo-rewards than the other. At the beginning of the task, the two options were equally rewarded, but the probabilities changed as the task unfolded, decreasing for the most selected option. The results indicated that people preferred the most pseudo-rewarding option, even when it yielded less final reward. In addition, activity in the VS correlates with not just reward prediction errors but also pseudo

reward prediction errors and with the preference for the most pseudo-rewarding option.

Introduction

Rewards are desired outcomes that have an effect on and motivate behaviours. Individuals analyse their reward history to determine and choose between alternative courses of action; therefore, the information provided by the incentives is crucial. The significance of prior knowledge suggests that our behaviour is influenced by the outcomes of our previous actions, whether they were done to obtain some benefit or to avoid punishment (Sutton and Barto, 1998; Shteingart and Loewenstei n, 2014). Individuals attempt to increase the frequency and intensity of activities that result in rewards for a variety of reasons, such as satisfying a need or avoiding undesirable consequences. In fact, the way in which feedback influences goal-directed behaviours results in behavioural changes and influences brain processing. Therefore, it is essential to comprehend how the history of rewards is encoded in the brain, leading to the formation of goal-directed behaviours-guiding learning habits (Lak et al., 2021; Lee et al., 2012; Niv, 2009). Reinforcement learning (RL) is a framework that upholds trial and error as a learning mechanism, bridging reward-based learning and decision-making. Most RL theories focus on reinforcing behaviours based on rewards prediction errors (RPEs), which seek to describe whether actual outcomes are better or worse than anticipated (Sutton and Barto, 1998).

O'Doherty et al. (2017) demonstrated that the computational principles underlying standard RL theories have extensive experimental correspondence in simple learning tasks but have limited efficacy in complex

settings (Botvinick, 2009; Ribas-Fernandes et al., 2019). The more steps that precede access to the reward, the more challenging it is to explain the behaviour (Botvinick, 2012; Ribas-Fernandes et al., 2011). Different accounts assume that individuals divide expansive routines into simpler behavioural units, thereby creating a hierarchical, progressive partition that seeks to obtain a reward. These behavioural dynamics form the basis of Hierarchical Reinforcement Learning (HRL), a computational algorithm that has evolved to address the difficulty of explaining extensive behavioural pathways (Botvinick, 2012; Hengst, 2012). As a result, HRL algorithms propose progressive learning in multiple stages, thereby simplifying decision-making and learning. This strategy reduces the complexity of the behaviours by arguing that the completion of each step brings the subject closer to the ultimate reward. However, the pursuit of completing these stages would be an objective itself, and as a result, it would generate expectations regarding its outcome (Botvinick et al., 2009). Therefore, the disparity between the anticipated and actual outcome would generate its own prediction error (Botvinick et al., 2012; Diuk et al., 2013b). If feedbacks facilitate learning, then this class of pseudo-feedback display in intermediate states would also be able to influence behaviour.

Therefore, RL is a form of adaptation in which an individual uses past experience to improve future decisions, and well-differentiated neural substrates have been reported to be activated during this process (Lee et al., 2012; Shteingart and Loewenstein, 2014; Holroyd et al., 2017). Extensive research indicates that the human ventral striatum is a crucial component of a circuit that is activated during learning, weighing the value of the current outcome in order to make better decisions in the future (O'doherty, 2004; Montague et al., 2006; Rangel et al., 2008). Indeed, it is well established that striatum-projected dopaminergic neurons play a crucial role in RPE

computation and encoding (Schultz et al., 1997; Delgado, 2007; Heekeren, 2007). Moreover, given that value functions can change in response to rewards and punishments, as well as pseudo-feedback processing, they might present distinct brain correlates. In fact, the striatal contribution extends to RPEs generated after receiving pseudo-feedback, indicating that the brain encodes parallel RPEs to adjust not only to the history of rewards but also to the pseudo-feedback evidence collected (Ribas-Fernandes et al., 2018). However, the involvement of these mechanisms is still not well understood. In the present study, we modified the experimental paradigm of Mas-Herrero et al. (2019) by presenting participants with a two-step task with varying probabilities of receiving a pseudo-reward but the same initial probability of final reward. Importantly, as the task progresses, the final reward of the most selected option will decrease, allowing the neural mechanisms underlying the preference for the most pseudo-rewarding option to be studied. We hypothesised that activation of the ventral striatum is essential for the parallel processing of reward prediction errors and pseudo-reward prediction errors, even during sub-optimal decisional behaviour.

Participants

The experiment included twenty right-handed people (M = 23.95 years old, SD = 4.13, 16 women). All participants provided written informed consent and were compensated at a rate of 10€ per hour. Two participants were dismissed one because the recordings were severely distorted and the other because he decided to withdraw from the experiment during the execution. The local ethical committee approved all procedures.

Task Design

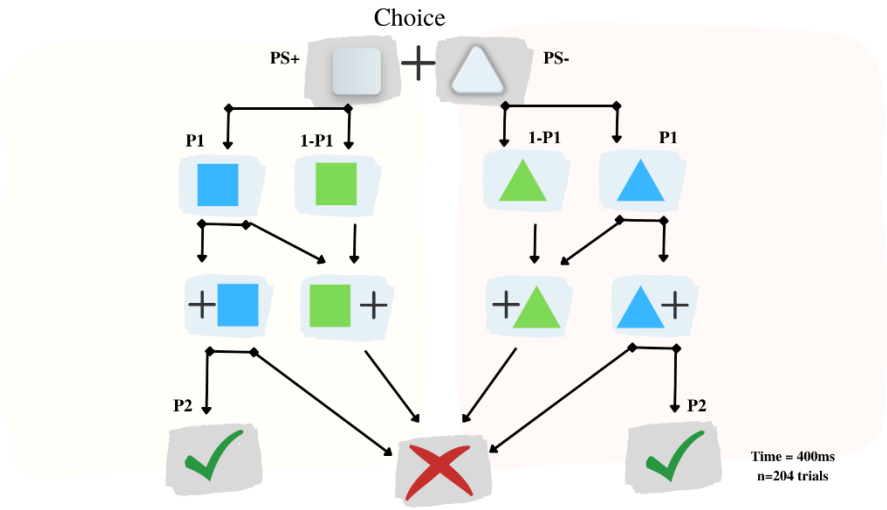


Figure 5.1. Decision-making task. The task involved two steps. In order to earn money in the subsequent step, it was necessary to first obtain a specific colour (orange in this case) (green tick). In free-choice trials, participants were required to choose between two options, one of which (PS+) had a higher probability of obtaining pseudo-rewards (p1PS+) but a lower probability of obtaining final rewards after positive pseudo-feedback (p2PS+). The other figure (PS-) displayed the opposite pattern. Initially, the two figures had the same probability of obtaining the final reward. However, as the task progressed, p1 changed, resulting in differences in the final reward probability (p1 x p2). In free-choice trials (one-third of the options), participants could choose between the two figures. In the remaining trials (two-thirds, forced-choice), only one figure was displayed.

A paradigm introduced by Mas Herrero et al. (2019) was adapted to the purpose of this research. This experimental design proposes a sequence of two hierarchical stages per trial, with the first step providing pseudo-feedback and the second step's result providing feedback to raise their final pot. The participants were instructed to get as many rewards as possible. Participants had to choose between two informative geometric cues within one second at the start of each trial by pressing right or left pads. Right after the selection, a fixation point was displayed for 1200 ms. Then,

either stimuli had a probability p1 of providing a positive pseudo-feedback, which was represented by a figure painted in blue or green presented in the centre of the screen for 1000 ms, followed by a fixation point jittered for 500 ms. Participants who got a certain colour (blue in half of the participants, green in the other half) could potentially win money at the next stage. They would certainly lose the trial if they were given the wrong colour (green). After that, the figure was displayed on the left or right side of the panel; then, the participant had to press the suitable button to indicate its position within 1000 ms. 1200 ms after the response, participants received the final feedback. After receiving positive pseudo-feedback (e.g., blue cue), the individual had a probability (p2) of receiving a reward (0.2 €, green tick) or no reward (1- p2, red cross). The participants would surely lose the trial every time they were given the wrong colour (e.g., green) at the first stage; therefore, a red cross was always presented in this situation. The outcome, either cross or stick, was presented in the centre of the screen for 1000ms. Right after, a fixation point was presented until the total trial lasted 11 seconds. Finally, a jitter of mean 1500 ms (randomly jittered between 500 and 2500 ms, with 400 ms increments) was introduced before the next trial.

In the beginning, the two cues had an equal probability of winning each task (p1 x p2). One Cue (PS+) was linked to a higher likelihood of receiving colour (p1(PS+) = 0.7) but a lower likelihood of receiving input after receiving positive pseudo-feedback (p2(PS+) = 0.3). The other cue (PS-) had a lower chance of obtaining positive pseudo-feedback (p1(PS-) = 0.3) but a higher chance of receiving feedback after receiving positive pseudo-feedback (p2(PS-) = 0.7). As a result, participants started the task invariantly with the same overall chance of achieving a final reward (p(reward)=.21). This paradigm consisted of 144 trials where participants were required to choose between the cues in 48 trials (free choice trials). In the remaining 96 trials

(forced trials), one cue was shown on one side of the board, and the subject had to choose it by clicking the appropriate button (48 trials for each shape). As a result, the participants had the opportunity to accumulate evidence of the probabilities for the two shapes without having to rely on their free-choice choices.

The task was changing its probability along with the trials regarding the participant's choices. After obtaining a final reward in the Free Choice trials, p1 of each cue changed, with the chosen form decreasing 0.02 and the non-chosen form increasing 0.02. This manipulation helped to determine to what extent participants would stick to one of the choices even though the final incentive likelihood (p1 x p2) was lower for a cue and higher for the other. This manipulation was not explained to the participants.

fMRI data acquisition

fMRI was performed on a 3-Tesla Siemens magnetic resonance scanner at The Mind, Brain and Behaviour Research Center in Granada, Spain. In order to acquire a T2*-weighted echo-planar imaging (EPI), a sequence was used, with a repetition time (TR) of 2000 ms (three runs of 275 scans each) and 35 descendent slices with a thickness of 3.5 mm^3 (echo time = 25 ms; flip angle = 80º; voxel size = 3.5x3.5x3.5 mm^3; matrix size = 68 x 68). Slices were aligned at 30º to intersect the anterior and posterior commissures due to control artefacts in the OFC and the anterior ventral striatum. Therefore, the fMRI was optimised to reduce susceptibility-induced BOLD sensitivity losses (Weiskopf et al., 2006). High resolution T1-weighted anatomical images were obtained after the functional task (192 slices; image matrix 256

x 256; voxel size = 1 mm3; repetition time = 2500 ms; echo time = 3.69 ms, flip angle = 7°). Participants' heads were padded inside the head coil to prevent unnecessary movement.

Temporal difference Model (TD0)

We used a temporal difference learning model (TD0) to estimate RPEs associated with feedback and pseudo-feedback. We fitted four different parameters: one that modulates the role of pr (PS), the other related to how random subjects performed, and two different learning rates: one for pseudo-feedback processing and the other for feedback processing. The negative log-likelihood was minimised to optimise the four parameters. We used temporal difference learning (TD0) to predict RPEs related to both feedback and pseudo-feedback. The action values were dependent on the history of pseudo-feedback and feedback. In this experiment, two RPEs were computed: when the pseudo-feedback was delivered (RPE1) and when the feedback was achieved (RPE2). The pseudo-feedback modulated RPE1, signalling whether the sub-goal had been met or not. RPE2 and the action values were determined in tandem by the final reward. The action value was updated when the pseudo-feedback was presented. The model included the parameter PS, which was calculated by minimising the negative log-likelihood using the fminunc function in MATLAB, modulating the weight of the pseudo-reward in an individual bias. In study 2 of the current thesis, it was determined that model 3 (see Chapter 4) accurately described the decision pattern throughout the hierarchical task; therefore, it was used to calculate RPE1 and RPE2 (see Study 2 for details on the model).

fMRI analyses

The SPM12 Matlab toolbox (https://www.fil.ion.ucl.ac.uk/spm/software/spm12/) was used to pre-process the whole-brain images. An image from averaging was generated after slice timing correction and realignment (descending order, sinc-interpolation, reference slice 17). The T1 image was co-registered with this mean image and then was segmented. Following that, data was normalised using fourth-degree B-spline interpolation. Finally, we smoothed the normalised volumes using an 8-mm 'full width at half maximum' kernel.

Regarding the preparation of the data, the presentation of the figures, the delivery of pseudo-feedback and the final reward were included in the model, as well as the choice of left or right on the keyboard to minimise directionality bias. To achieve a regular scale of the output regression parameters, the parametric regressors of RPE1 and RPE2 were standardised using a mean of 0 and a standard deviation of 1 (Erdeniz et al., 2013; Mas-Herrero et al., 2019) and were entered as covariates in a first-level analysis. A 128 s high pass filter was applied to the time series.

Two contrasts were proposed for the first-level analysis, one for RPE1 and the other for RPE2. Using one-sample t-tests, these contrast images were entered into independent second-level group analyses. Two analyses were performed, a whole-brain analysis at a voxel level threshold of $p < .001$ and an ROI at VS based on a bilateral Nucleus Accumbens mask extracted from Hammers' probabilistic atlas ($p < .05$).

Results

Behavioural results

Participants had a consistent preference for the Cue 1 along the experiment (M = 0.57, STD = .09), which was significantly above the randomness scenario of 0.50 ($t(17) = 4.3$, $p < 0.001$). Thus, we conducted individual binomial tests for preference for PS +, finding that 7 of 19 subjects showed a significantly higher preference ($p < 0.05$) for Cue 1. In comparison, only one individual showed a significantly higher preference for Cue 2. The preference for Cue 1 validates the influence of pseudo rewards when choosing an option that delivers positive pseudo-feedback more often. On average, participants won in one of every seven trials (M = 20.1, STD = 3.7), which meant an additional monetary compensation of €4.02.

fMRI results

Table 5.1. Brain Activity correlated with RPEs

Region	# Voxels	T	x	y	Z
Pseudo-Feedback					
Right Ventral Striatum	227	6,298	14	14	-4
Left Ventral Striatum	102	4,953	-8	12	-4
Feedback					
Ventral Striatum	385				
Right Ventral Striatum		7,018	8	16	-6
Left Ventral Striatum		7,933	-6	8	-6
Left Superior Frontal Gyrus	1073	6,454	-18	38	44
Right Superior Frontal Gyrus	209	6,243	22	38	44

*All the correlations between brain activity and the RL model with a significance of p (uncorrected) < 0.001 at a peak-voxel level.

The table 5.1 shows the results for the whole brain analysis for RPE1 and RPE2. We found bilateral ventral striatum activation for the RPEs after receiving both pseudo-feedback (RPE1, $p < .001$ uncorrected; right: x = 14, y = 14, z =-4; le ft: x = -8, y = 12, z = -4) and the final feedback (RPE2, $p < .001$ uncorrected; right: x = 8, y = 16, z = -6; left: x = -6, y = 8, z = -6). Also, we found an activation of the SFG in the left and right hemisphere, but just in the case of the feedback (RPE2, $p < .001$ uncorrected; left: x = -18, y = 38, z = 44; right: x = 22, y = 38, z = 44). All of these clusters were FWE corrected at a cluster-level ($p<0.05$), except the left VS in the RPE1 contrast.

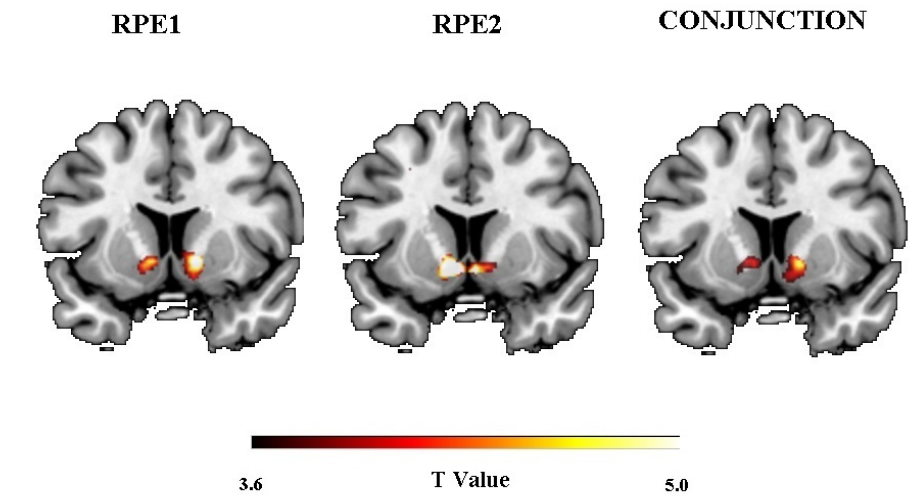


Figure 5.2. Nucleus Accumbens Activity of RPEs. Activation of NAcc RPE ($p < 0.05$ FWE correction) based on an anatomical mask of the bilateral Nucleus Accumbens (Hammers et al., 2003) for RPE1 (left), RPE2 (mid) and the conjunction of the two contrasts (right).

Figure 5.2 shows the brain activity in the ventral striatum based on the anatomical mask of the bilateral NAcc related to RPEs ($p < 0.05$ FWE correction). More specifically, the T-maps depict the brain regions where a positive relationship between BOLD activity and the RPE1

(pseudo-feedback) and RPE2 (feedback) was observed across trials. As displayed, the last panel on the right depicts the conjunction of the two preceding relationships (RPE1 and RPE2), and exposed a common activation in both the left and right ventral striatum ($p < .05$ FWE; right: $x = 14$, $y = 10$, $z = -4$; left: $x = -8$, $y = 10$, $z = -4$).

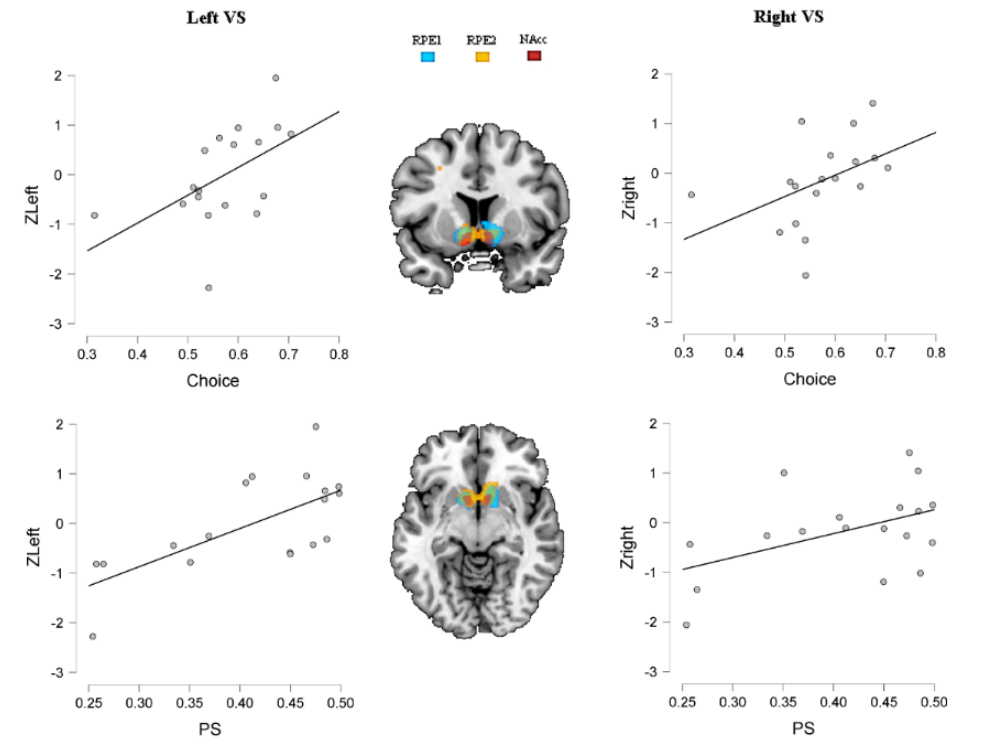


Figure 5.3. Brain-behaviour relationship. Scatter plots depicting the correlations between NAcc activity (delimited using the Hammers et al. 2003 atlas) and the preference for the most pseudo-rewarding figure (top) and the Ps parameter of the RL model (bottom) for the left and right areas. In the centre, the T-maps of whole-brain activity at $p < 0.001$ (uncorrected) for RPE1 (blue), RPE2 (orange) and the NAcc mask (red) are shown for visualisation purposes.

The figure 5.3 shows the correlation between the activation of the NAcc defined by the Hammers et al. (2003) atlas and the Ps parameter of the RL model. Results showed a significant relationship with PS parameter in both,

the left VS (r (17) = 0.57, p = .01), and the right VS (r (17) = 0.49, p = .03). In addition, the other parameters of the RL model did not show a significant correlation with the striatal representation. Moreover, we also assessed such VS sensitivity in relation to the proportion of the selection of the option that delivered pseudo-rewards more often. We found a correlation in the left VS (r (17) = .51, p = .02) and a marginal correlation for the right VS (r (17) = 0.42, p = 0.073).

Discussion.

The aim of this study was to determine the neural substrates of the processing of pseudo-rewards when they generate a preference towards them, even when this bias implies a cost and reduces the possibility of obtaining rewards. To test these predictions, we used a learning task whose original version was designed by Mas-Herrero (2019), which is composed of two levels. Participants needed to obtain positive pseudo-feedback to have the possibility of accessing a monetary reward. Participants developed a preference for the option that provided pseudo-rewards more frequently, as has been identified in other investigations (Ribas-fernandes et al., 2019; Mas-Herrero et al., 2019). However, each time this option was chosen in the present study, the option that delivered fewer pseudo rewards increased its chances of delivering the final reward. Thus, such imbalance implied that the participants maintained their preference for engaging in sub-optimal behaviour. We demonstrated parallel processing of prediction errors as indicated in the literature (Schultz, 1997), and we evidenced its correlation with the activation of the ventral striatum (VS). Moreover, RPEs produced at

both moments, after receiving pseudo-feedback and after the final feedback, have a representation in VS. At the same time, it was evidenced that the activation of this area is related to the decisions and the value given to the pseudo rewards by each participant.

Results of the present study show that the VS encodes both RPE1 and RPE2. These results support the notion that PEs related to sub-goals and final goals involve the same brain correlates (Mas-Herrero et al., 2019; Ribas-Fernandes et al., 2011, 2019; Diuk et al., 2013b; Shahnazian et al., 2018; 2019). In addition, they emphasise the role of the striatum in calculating prediction errors and learning the value of stimuli (Peters et al., 2021; Marche et al., 2017; Ma et al., 2014), as well as the fact that dissociable PEs are incorporated within the same behavioural transition (Daw et al., 2011; Diuk et al., 2013b). Different studies have previously investigated the relationship between hierarchical RL and decision-making with different levels of complexity (Doya, 1999; Dezfouli and Balleine, 2012; Diuk et al., 2013b). We adopted a standard TD0 approach similar to Diuk et al. (2013b) and Mas-Herrero (2019), where every bit of information delivered between the initial decision and the outcome could result in an RPE, therefore used to update reward expectations. However, in our case, RPE1 and RPE2 are mediated by different learning rates (α_1 and α_2), which would indicate how individuals learn at a different speed and if they process RPEs in a particular way. Moreover, an additional parameter weighted the value of pseudo-feedback in an individual bias.

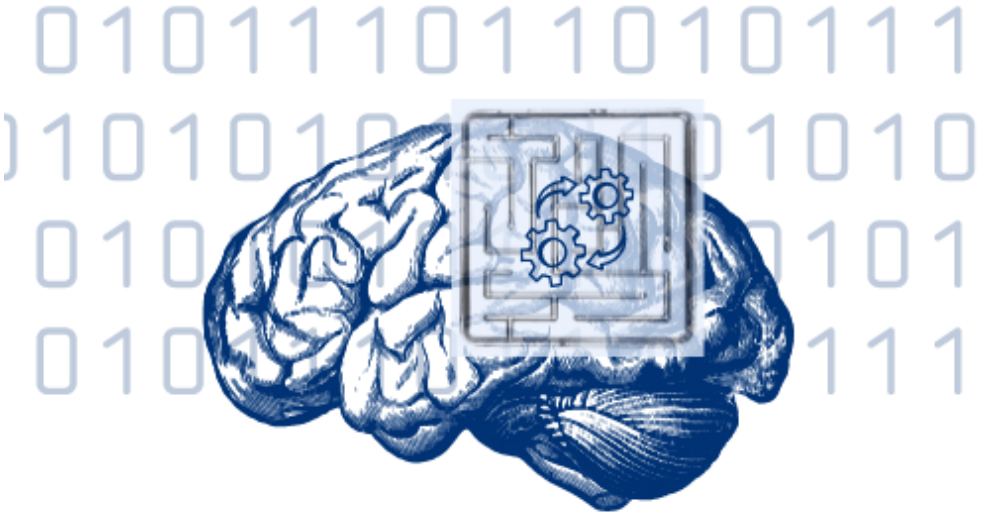
Furthermore, we show that the behaviour of individuals is related to the BOLD signal identified in the VS. In fact, a higher striatal sensitivity to RPEs was related to a higher choice of the most pseudo-rewarding option. This finding validates and extends the results provided by Mas-Herrero et al.

(2019) and Daw et al. (2011), where a model-free determined the elections and their representation in VS as opposed to a model-based. Our findings imply that the ventral striatum stores learning signals and that these learning signals play an essential role in decision-making. However, the possibility that the striatum can store numerous concurrent predictive error messages raises the question of how these learning signals differ. This critical issue, which was previously introduced by Diuk et al. (2013b) and Daw et al. (2011), will require more future research. Rather than having a role in learning per se, one idea is that the striatum is implicated in integrating predictive error signals to direct behaviour. This reasoning is consistent with previous research indicating that the ventral striatum is not only crucial for recording reward prediction errors but also for determining the behavioural relevance of events (Klein Flugge et al., 2011; Iglesias et al. al., 2013; De Lange et al., 2018). For example, Klein-Flugge et al. (2011) found that the ventral striatum responds predominantly to events relevant to behavioural orientation, in contrast to midbrain activity that represents RPE signals in the absence of a behavioural policy. In fact, in our research, the representation of RPE1 in brain maps was concentrated in VS, while RPE2 was also related to other areas, in concrete, the Superior Frontal Gyrus. Furthermore, some authors hypothesised that striatal activity might influence feedback processing, thus increasing the relevance of those events and their related behaviours for future decisions, particularly when an event gains special attention (Den Ouden et al., 2012; Berridge, 2007; Friston, 2009).

In this study, Cue PS+ provided more pseudo-rewards which could be interpreted as additional information connected with the monetary reward. This imbalance potentially had a motivating influence on their selection. It can be argued that there is a preference for the source that provides additional information on the reward as a result of an observational effect

(Beierholm and Dayan, 2011). This preference for positive pseudo-feedback relegated the feedback processing of the final reward, which was more favourable for the other option in most of the participants. Therefore, this study shows a complete behavioural trajectory guided by pseudo-feedback, despite leading to sub-optimal behaviour, which may be due to participants not paying attention to the value of the conditioned reinforcer without taking into account the stimulus that predicts the lack of reinforcement (Zentall, 2016).

In conclusion, the findings on the relationship between prediction errors and VS activation indicate the potential of pseudo-feedback as a learning accelerator and its effect on decision-making. In addition, these brain mechanisms would be present in sub-optimal situations in which individuals would remain in a lost pattern, demonstrating that the propensity for pursuing sub-goals can result in undesirable behaviour.



Chapter 6. Discussion

Chapter VI: General Discussion.

Introduction.

Three studies were conducted in this thesis to investigate the behavioural, computational and neural correlates of sub-goal-oriented behaviour during sub-optimal decision making. On the one hand, we researched the behavioural substrates of sub-goals as a driver of learning, even when their guidance can lead to maladaptive choice behaviour. On the other side, we investigated the brain correlates to pseudo-feedback processing in Studies 2 and 3, diving into the oscillatory activity and the BOLD signals underpinning the Pseudo Reward Prediction Errors' encoding. Summarising and over-viewing the individual findings from each research, we will revise the major findings from the studies that compose this thesis. By combining them into a more holistic picture, we will be able to discuss this evidence at the light of the research aims and existing literature. Nevertheless, each study's related section has a more extensive description of the findings (chapters 3, 4 and 5).

Summary of the findings

The purpose of the **study 1** was to determine if pseudo-rewards may influence choices to induce sub-optimal behaviours. By using a series of a modified version of a two-step hierarchical task, we established that individuals favour the most pseudo-rewarding alternatives, even if they are

not the most advantageous in terms of ultimate rewards. These findings demonstrate that intermediate information on decision-making is self-reinforcing and may result in sub-optimal choice behaviours, at least in a hierarchical setting. Interestingly, we observed a preference for the more pseudo-rewarding option throughout all of our tests, both when the two alternatives resulted in the same final reward (Experiment 1) and when choosing this figure resulted in a decrease in the probability of obtaining the final reward (Experiment 2). Tasks with secondary goals have previously highlighted the relevance of information on the different stages necessary to accomplish a task or achieve a final reward (Ribas-Fernandes et al., 2011).

The purpose of the **study 2** was to determine the influence of pseudo-feedback on biased decision-making and its associated brain oscillatory activity. For this purpose, different temporal difference models were presented to explain participants' behaviour and their association with oscillatory theta activity, which is related to prediction error encoding. The model best adjusted to the behavioural evidence was the one that proposed two different learning processes, one for each stage of the task. In addition, theta oscillatory activity increased after getting pseudo-feedback and feedback. These findings of theta oscillations on the mid-frontal- cortex are aligned with previous findings of ACC activity and prediction error encoding (Holroyd and Yeung, 2012; Gruber et al., 2013; Mas-Herrero and Pallarés, 2016; Shahnazian et al., 2018). Also, as in this study 2, sub-goals are supposed to allow improved prediction of pre-states that lead to ultimate rewards, even with two independent systems running in parallel, as in Diuk et al. (2013b), but with independent Learning Rates per task level.

The goal of the **study 3** was to identify the neural correlates of processing pseudo-rewards that induce a preference for them, even though this bias

diminishes the likelihood of receiving rewards. The study's key finding was the statistically substantial connection between BOLD signals in the Ventral Striatum (VS) and prediction errors. Consequently, VS activations are implicated in both action levels; therefore, we found that predicted values of distinct forms of feedback are encoded on a similar scale and activated in the VS. Furthermore, these PRPEs engage the reward network in similar ways to RPE (Diuk et al. 2013b, Mas-Herrero et al. 2019), supporting their reinforcing properties which would justify their engaging nature which could lead to such sub-optimal behaviour.

The Reinforcing potential of pseudo-feedback.

The relevance of goal-related information about the stages required to accomplish a task or achieve a final reward has already been demonstrated (Ribas-Fernandes et al., 2018; Lou et al., 2021). The sub-goals serve a major purpose and are the key to sub-goal-directed behaviour, as it is a behavioural strategy to lower the task's overall complexity and direct attention and resources to a simpler set of subroutines (Mas-Herrero et al., 2019). These sub-goals enable a more accurate prediction of the pre-states that precede a final reward, as it has been extensively presented by Botvnick and collaborators (2011;2013). Particularly the experimental design proposed by Diuk et al. (2013b) is comparable to the two-step task employed in this thesis ever since it employed a parallel prediction error computation at different levels. Notably, it has been demonstrated that the prediction errors associated with these pseudo-feedbacks engage areas of the reward network, such as the ventral striatum (VS; Diuk et al. 2013b, Mas-Herrero et al., 2019), implying

that their reinforcing properties justify their engaging nature, which may result in such sub-optimal behaviour. Mas-Herrero et al. (2019) discovered a correlation between Ventral Striatum activity and preference for positive pseudo-feedback, indicating that individual variance in terms of preference may be connected to the influence of pseudo-rewards on the reward network and our goal-directed behaviour. So, first of all, it is clear that the assumption of brain areas involved in the reward circuitry could justify the fact that sub-goals have a reinforcing potential in choice behaviour. Particularly in the study 1 / experiment 1, this behavioural preference for pursuing pseudo-feedback was stressed and validated.

Therefore, the hierarchical organisation of behaviour might imply a relationship between goal-directed behaviour and other systems which aim to decompose task units in simpler settings. For the purpose of the discussion, we could name it sub-goal-directed behaviour as a subordinate concept to the mainstream theory of goal-directed behaviour (Polania et al., 2014). Still, the reinforcing nature of pseudo-feedback had not been tested in the light of risking optimality, particularly within an RL approach.

Pseudo-feedbacks as drivers of sub-optimal choice behaviour

While several theories show that people are directed by statistically optimum decision-making (Bogacz, 2007), the results of experiment 2 of study 1 suggest that information about intermediate stages may influence judgments toward poor choices. These findings corroborate previous research demonstrating sub-optimal choice behaviours or maladaptive decision-making, where learning from the information available in the

environment might mislead individuals to limit their access to an ultimate reward (Stagner and Zentall, 2010, Chow et al., 2017; Colaizzi et al., 2019). In a more behaviourist approach to the learning theories, agents might be biased when exposed to more salient stimuli or when confronted with decisions with a higher predictive value that results in a predilection for these inter-linked cues that guide one's actions (Colaizzi et al., 2019). Thus, the sub-optimal choice behaviours indicate that pseudo-feedbacks act as a probable discriminative stimulus, which provides information with a positive predictive value (McDevitt et al., 2016; Sears et al., 2022). The sub-optimal imbalance between options in respect to the point of subjective equivalence (PES) remarks this bias toward pseudo-rewards.

Adding an additional layer to the discussion, it has been reported that individuals exhibit a preference towards stimuli that have previously related to a reward; therefore, they trust such signals to guide their behaviour. This approach has been coined as the well-known sign-tracking effect (Amaya et al., 2020; Lesaint et al., 2015; Meyer et al., 2012). Results of the current thesis show that, given that pseudo-feedbacks are perceived as required informative cues, sub-goals play a significant role in generating sub-optimal preferences. This suggests that when any information linked to an ultimate response is assumed as a mandatory announcement of a stimuli appearance, individuals search for these signs and avoid other alternatives that might not contain the informational cue. In this view, pseudo-feedbacks would act as the conditioned stimulus in sign-tracking techniques, becoming appealing and inducing approaching behaviours toward them. For instance, various animal (Bromberg-Martin and Hikosaka, 2009; Anderson et al., 2015) and human (Eliaz and Schotter, 2010) studies have suggested that information is intrinsically rewarding and that certain biases or even sub-optimal behaviours may be explained by a preference for informative over

non-informative stimuli. However, while one might argue that participants preferred a more pseudo-rewarding option because it was more informative, in our experimental designs, the available options delivered the same amount of information, or at least with the same frequency. Another possible interpretation of these sub-optimal biases would be on the bases of the temporal delay between feedback and pseudo-feedbacks. Still, the role of uncertainty is unclear because, in all the paradigms, forced trials were included to unfold all the information available. Additionally, the fact that the time required to achieve the ultimate reward was the same in both alternatives ruled out the idea of a preference for one option owing to temporal delay discounting (Kirby et al. 1999). Finally, participants' preferences cannot be attributed to inconsistencies in the stimuli's presentation since they were equally exposed to the different options, and these features were counterbalanced across participants (Niv et al., 2002).

Is a sub-goal-directed behaviour a trade-off along with habitual behaviours?

Results of study 1 show that sub-goals have an impact on learning, perhaps because of their capacity to orient fewer demanding actions into a larger routine towards a goal, and that pseudo-feedback pursuance can derive from motivated behaviours towards their acquisition. Considering that individuals seek to maximise their chances of getting the essential resources to survive, it is unclear why sub-goal-directed behaviours can land into sub-optimal choice behaviour. However, it is important to note that individuals have a variety of techniques to manage ambiguity, including

gathering information from prior encounters with the environment in order to create knowledge about their odds of accessing the resources. This exchange enables individuals to forecast and assess new data to improve their learning and their capacity to deal with uncertainty. As a result, several mechanisms aid in monitoring the outcomes of decisions, those that primarily respond to the characteristics of the stimuli and those that anticipate future actions.

Two primary strategies for behaviour control have been identified and coined as habitual behaviour and goal-directed behaviour. In the first case, actions are guided by an unanticipated external stimulation (stimulus-driven), whereas goal-directed behaviours appear to be motivated by a specific reward that has been spotted and accounted as a probable outcome (Nachum et al., 2018). These two, apparently, separate types of interaction with the environment and value-based learning have the potential to impact individuals' decisions in different ways. One potential source of sub-optimality might be those behaviours that have been referred to be habitual because they lack the inherent monitoring and error detection mechanisms seen in goal-driven behaviours (Decker et al., 2016).

People's reactions, according to this idea, are strongly dependent on the amplitude and salience of a stimulus, and this could partially explain the attraction toward positive pseudo-feedback instead of focusing on the final reward, as has been stated in all the studies. Therefore, in our experimental setting in study 1, pseudo-feedback would indicate a certain probability of getting a monetary incentive, but still, participants would keep their preference to pursue this pseudo-reward rather than focus on the most optimal strategy to get a final beneficial outcome. Perhaps the pseudo-feedback's salience was enough to capture the individual's attention in a look-alike habitual behaviourist setting. As stated above, similar

behavioural responses can be observed in sign-tracking, where individuals guide their behaviour based on external influences (stimuli features) while assuming a reactive position. Decision-making occurs more swiftly and automatically in a stimulus-driven behaviour, which is closer to basic adaptive processes. As a result, the interaction with the environment is more reactive, making it harder to employ in more complicated behaviours where the stimuli's salience is not the primary element.

Yet, the argument of habitual learning as a central explanation for sub-optimality is not enough to describe the evidence of learning and the proactive behaviour towards pseudo-feedback found in Studies 1, 2 and 3. Therefore, we have proven that sub-goals and their respective pseudo-reward prediction errors are encoded similarly and in parallel to the ultimate rewards, which could indicate a somewhat mediate stage between habitual and goal-directed behaviours. Before digging further into a more computational modelling discussion and its place in the RL framework, it is crucial to observe that a goal-directed process assesses various actions and their predicted consequences, whereas a quicker and more automatic habitual process connects rewards with signals, enabling repetition of previously successful activities (O'doherty et al., 2017).

However, even when traditionally habitual and goal-directed behaviours have been assumed as independent monitoring processes underlying decisions, their interaction in learning has been demonstrated in rats (Dickinson et al., 2002; Chow et al., 2017) and humans (Valentin et al., 2007), where both strategies (habitual and goal-directed) appear to act in combination and coexist. Indeed, their concurrent existence in learning and decision-making may indicate that they work in tandem to promote resource access and survival demands. A point of convergence between these two

behavioural approaches may help, to some extent, to explain why pseudo-feedback leads individuals into sub-optimal habits and decision-making biases.

Indeed, not just habitual learning but also goal-directed learning has its own computational representation within the RL theories. For instance, all the experimental designs that took place in this thesis included the principle of progressive learning to favour an individual's search for rewards, yet the evidence of sub-optimal choice behaviour was consistent with the different studies. These findings of a persistent bias have been challenging since learning suppose to happen in an incremental manner, and in principle, hierarchical tasks (two-step tasks) should be able to facilitate the learning of optimal behavioural strategies.

Oscillatory activity and pseudo-feedback processing.

One of the main findings of the second study was the presence of theta activity in all conditions suggesting that similar brain systems are recruited during both pseudo-feedback and feedback. Midfrontal theta activity has been associated with activation of the anterior cingulate cortex (ACC) (Holroyd and Yeung, 2012; Gruber et al., 2013; Mas-Herrero and Pallarés, 2016; Shahnazian et al., 2018). Furthermore, assuming that common brain processes are used to calculate prediction errors at both action levels, predicted values for various forms of feedback are encoded on a similar scale and activate common brain regions (Diuk et al., 2013b; Levy and Glimcher, 2012). This would also be compatible with earlier results about the correlations between hierarchical temporal abstraction and the mPFC

(Balaguer et al., 2016; Holroyd et al., 2018; Ribas-Fernandes et al., 2019). It is important to note, however, that oscillatory activity was greatly increased while processing feedback linked with the most pseudo-rewarding option. Thus, the apparent preference for the most pseudo-rewarding figure implies a sub-optimal bias, considering that the most optimal option was predefined for the least pseudo-rewarding figure, yet the most beneficial.

Additionally, there were substantial changes in the elicited mid-frontal power for theta band. The variations in theta power increase found in this study were consistent with earlier findings regarding gain-loss comparisons (Sambrook and Goslin, 2015) and expected and unexpected reward dissociation (Paul et al., 2020). While there is a significant difference in gain between the available options, gain from PS+ was always more surprising, therefore generating a conventional reaction to novelty, similar to feedback-related negativity (FRN). In normal settings, participants would gradually redirect themselves to the most ideal choice (Balaguer et al., 2016), but this does not seem to be the case in any of the three studies. It has been suggested that frontal-midline theta oscillations (4-8 Hz) are a complementary correlate of the event-related error monitoring process (Hajihosseini et al., 2013), linking RPE signals with cognitive control implementation (Cavanagh and Frank, 2014; Holroyd and Umemoto, 2016). Increased frontal-medial theta power occurs during error processing, as well as in response to temporal differences and unexpected events during reward processing (Gheza et al., 2019).

Cavanagh and Frank (2014) proposed that prediction errors in an RL environment are associated with behavioural adaptation and learning (Cavanagh et al., 2010), probably through signalling the requirement for greater cognitive control in response to prediction error updating. Indeed,

theta activity demonstrated a stronger behavioural link with changes in prediction error updating and reward processing. Additionally, theta activity may be capable of capturing the neurological impacts of long-distance connections between the medial and lateral prefrontal cortex and the striatum (Smith et al., 2015; Gheza et al., 2019). By exploiting these properties, assessing the induced front-medial theta rhythms throughout varied HRL contexts provides unique insights into neural mechanisms underlying pseudo-reward-based learning and decision making.

Additionally, theta oscillatory activity and its association with PRPE encoding, as identified in study 2, confirmed a neural correlate in the midfrontal cortex during sub-optimal decision making, implying the necessity to broaden the scientific ground regarding theta oscillations during RL. Indeed, the association between prediction errors and theta activity in the midfrontal cortex verifies the pseudo-feedback scope as a predictor of learning not only in conventional decision-making schemes but also in sub-optimal settings in which individuals remain in a losing pattern.

The Ventral Striatum in an Actor/Critic architecture during sub-optimal choice behaviour.

The findings in study 3 suggest that the VS encodes in parallel, both the PE coming from pseudo-feedback (PRPE) and reward processing (RPE). fMRI results support the idea that PEs for sub-goals and final goals are processed in the same areas (Mas-Herrero et al., 2019; Ribas-Fernandes et al., 2011, 2019; Diuk et al., 2013b; Shahnazian et al., 2018; 2019). Also, as a result of the study 3 findings, the Ventral Striatum (VS) representations are

suggested as a predictor of learning during sub-optimal choices. In sub-optimal conditions, when people continue to lose, these brain processes would be still present as it happens in optimal choice behaviour, where the behavioural strategies are adjusted to avoid biases. An approach that might provide insights to interpret current results is the actor-critic architecture of RL models. This architecture, which has been widely used to characterise decision-making and RL's brain processes, establishes a differentiation between learning about states and actions (Cohen and Frank, 2009; Niv, 2009; Liakoni et al., 2022). To clarify, and as detailed in the introduction, the *actor* carries out the action, and the *critic* provides feedback on the actor's effectiveness while performing a task. A *critic* learns and makes predictions about the value of states and computes prediction errors when the agent transits between states. Therefore, the prediction errors are used to update not just the critic's *state-values* but also the actor's policy (Liakoni et al., 2022), that is, the function that maps states to actions. As a result, the actor selects actions in accordance with its learnt policy. The interest in the actor-critic model's relevance for understanding neural RL stems from the hypothesis that the differentiation between the ventral and dorsal striatum may widely transfer onto the actor-critic model's actor-critic distinction (Averbeck and O'Doherty, 2022; Van der meer and Redish, 2011). The *actor*, which has been reported to correlate with activity in the dorsal striatum and dorsolateral prefrontal cortex (dPFC), links sensory inputs to motor outputs (Holroyd et al., 2011; Van der Meer and Redish, 2011; Liakoni et al., 2022). Simultaneously, the *critic* component of a computational model, which is implemented by the orbitofrontal cortex (OFC) and ventral striatum (VS), gets sensory input to weigh the value of reinforcers. In this way, the critic gathers information regarding reinforcements and computes the value of ongoing events via temporal difference mechanisms (Padoa-Schioppa, 2011),

as well as reward prediction error signals indicating whether the values were worse or better than expected (Holroyd et al., 2016; 2017).

These signals are then transmitted to the dopamine system's neural targets (Ventral Tegmental Area) as phasic increases or decreases in dopamine activity (Schultz, 2013), where they are used by the actor to improve performance on the task and by the critic to optimise its predictions of future reward (Sutton and Barto, 1998). Therefore, the results of study 3 would support a key role of VS as the *critic* not only of final outcomes (as traditionally reported) but also for the intermediate steps needed to complete a task. In addition, this role could be on the bases of the maintenance of sub-optimal preferences. Interestingly, in study 3, we found BOLD signals in the VS related to individual behaviours. Striatal sensitivity to PRPEs was linked to a preference for the most pseudo-rewarding option but was also activated while processing the ultimate reward. This finding might indicate that the VS has a role as a critic and influence the actor's task performance during RL.

This intervention of the critic during intermediate states or pre-states is an evident indicator of disturbance of the conventional optimal action-selection process. Such a prominent role of the VS in the Actor/Critic architecture could be in line with the proposal of Niv et al. (2015), where a clear selection bias for some choices might contribute to decaying other alternatives while stopping exploration of other options. As hypothesised in this thesis, participants would prioritise the most reward-informative cues with a higher pseudo-rewarding value and target its features (i.e., colour, shape), implying that participants were learning in parallel about all levels of the selected stimuli but having a greater impact on the reward prediction mechanisms for those selected more frequently. Indeed, dopaminergic

prediction errors may have had an effect on the strengthening or weakening of corticostriatal connections encoding the characteristics of the regularly selected stimulus at the expense of other possibilities. In general, as the preference for particular options increased, the learning associated with unchosen stimuli began to decline consistently, as in a passive forgetting process (Beierholm and Dayan, 2010).

However, previous studies have suggested that not only the VS might play a role in this processing, and different proposals have been made to understand the brain processes underlying hierarchical reinforcement learning (HRL; Botvinick et al., 2009; Ribas-Fernandez et al., 2011). Holroyd et al. (2017) have proposed that ACC could be in charge of selecting the action policy, sometimes known as task or option. These authors propose that different options might be geared toward a certain goal and perform adjustments after receiving dopaminergic projections from the midbrain. In other words, the ACC would determine what to do; then, the actor would do the task, while the critic would monitor the result of these actions. So the options would be selected after being judged by RPE signals transmitted from the critic to the ACC via the midbrain dopamine system (Holroyd et al., 2017; Frank and Badre, 2012). This would be in line with the fronto-central theta activity found in study 2, which tracks PEs and has been related to activity in ACC and pre-SMA, among others (see, e.g., Mas-Herrero et al. 2016). In addition, this interpretation would explain why ACC was not found as be related to RPE or PRPE in study 3.

The assumption of the ACC as a mechanism for option selection may imply that the system relinquishes some control when the actor develops an effective action performance toward a goal, particularly when the amount of effort required to maximise reward is uncertain. It is possible that the

enticing nature of pseudo-rewards creates a mental state in which ultimate rewards are attained by achieving predefined sub-goals. In any case, behaviours with a high immediate reward value may be carried out independently of the primary ACC's goal of selecting the optimal option (Holec et al., 2014; Holroyd et al., 2017). Perhaps the excessive value ascribed to positive pseudo-feedback constrains the typical function of ACC in selecting the best alternative, particularly when a pseudo-reward is a necessary pre-state for an ultimate reward.

Indeed, considering the neural basis of the reward networks (Niv et al., 2015), it is conceivable that corticostriatal estimates are moulded to incorporate sensory factors that are thought to be important to the task (i.e., dimensions, levels, degrees, etc.; Bar-Gad et al., 2000), for instance, through selective attention processes (Corbetta and Shulman, 2002), where the actor and the critic hierarchical representations prioritise positive pseudo-feedback. Therefore, striatal circuits may also help to emphasise some inputs while ignoring others, which would mislead attention and generate potential biases (Frank and Badre, 2012; Pezzulo et al., 2018). Such attention filters, in turn, might be constantly updated based on the results of current choices instead of on the options selected fewer times (Canas and Jones, 2010), therefore providing a causal relation between the representation learning assumed by the individual and the RL model that describes behaviour.

Hierarchical Reinforcement Learning (HRL) to cope with uncertainty.

While a traditional reinforcement learning agent selects between concrete and straightforward actions, in HRL, agents divide into sub-routines, each with its own set of associated rules, that are utilised to accomplish sub-goals. Thus, with the use of Temporal Difference (TD) mechanisms and Reward Prediction Error (RPE) encoding, the agent can determine which subroutines are appropriate for each sub-goal (Hengst, 2012). Notably, relevant sub-goals are often not connected with primary reward, in contrast to overarching goals in a hierarchy. This exacerbates the difficulty of sub-goal setting and its prevalence in TD processes, which is perhaps the most difficult aspect of HRL (Botvinick, 2012; Wiering and van Otterlo, 2012).

Beyond the description of the sub-goals, there are predicted outcomes for each sub-task and therefore dubbed pseudo-reward prediction errors are created. Indeed, the two-step task with a monetary incentive used in the present thesis assumes that behaviour is hierarchically organised to reduce ambiguity. In contrast, such reliability and liability in TD mechanisms underlying RPE and PRPE might be the driving force towards biased choice behaviour. One of the purposes of study 2 was to determine the influence of pseudo-feedback on biased decision-making while using an HRL model. Therefore, a TD model was presented to explain the participants' behavioural inclinations, which are based on RPE and PRPE encoding. We developed and evaluated temporal difference models (TD0) with varied parameter fits follow-up the action selection and learning at different action levels. The model that was most closely aligned with behavioural data (model 3) advocated two distinct learning processes, one for each step of the task.

Additionally, the model that best explained the preference for PS+ incorporated two distinct learning rates (LR): one for updating values at the pseudo-feedback level and another for reward.

Additionally, model 3 introduced a third parameter that weighted positive pseudo-feedback in addition to the projected monetary incentive. This is, as long as a final reward had a monetary specific value (€0.2) and was included in the TD models as 1, we reckoned to find the best suitable value to represent pseudo-feedback in the model, which was 0.47, indicating a weight somewhere less than 1. This is fundamentally different to previous computational models used in HRL, such as in Diuk et al. (2013b) and Mas-Herrero et al. (2019). Due to the fact that it is not entirely based on actual rewards and that there is a persistent participant's decision bias, learning an optimal strategy becomes more difficult. Yet, in some circumstances, using pseudo-feedback may be useful due to its ability to accelerate learning and engage decision-makers. Additionally, the computation of two prediction errors demonstrates the HRL's ability to describe decisional paradigms with protracted phases of varying complexity. This also indicates that pseudo-feedback has a significant effect on learning and action selection. Similarly, as Botvinick (2012), Ribas-Fernandes (2017), and Mas-Herrero (2019) have demonstrated beforehand, the combination of stimuli in a chain of activities can have a distinct effect on learning.

Algorithmic constraints of HRL models and inefficient learning.

As stated above, the model that best matched the behavioural preference of the participants in study 2 was the model using two different learning rates

(one to update the values at the pseudo-feedback level and another to update the values at the reward level). This model concatenated TD0 independent designs for every task level, suggesting a hybrid approach under the assumption that parallel prediction error encoding might require independent learning rates. The feasibility of such an approach might bring back the typical constraints of flat RL algorithms in terms of dimensionality and scalability. Thus, the way of representing a hierarchical task decomposition and the complex nature of intermediate states between action and rewards might generate inefficient learning (Eckstein and Collins, 2020). Therefore, the critical point is to discuss both the limitations of the algorithms themselves but also the inefficient learning process due to the extent of behavioural routines.

Humans possess an enthralling capacity to accomplish goals in a complicated and continuously changing world, still outperforming sophisticated algorithms in terms of consistency and speed of learning. It is widely agreed that a necessary component of this skill is the use of abstractions and hierarchical representations that make use of the environment's structure to facilitate learning and decision-making. Nonetheless, little is known about how humans develop and then use those hierarchical representations (Eckstein and Collins, 2020). According to cognitive psychology and neuroscience, human and animal behaviour are hierarchically organised (Botvinick, 2012). Nevertheless, depending on the architecture to be used, designing a reasonable hierarchy requires domain knowledge and precise engineering in order for a method to appropriately describe sophisticated behaviours.

Imagine that your partner just baked an extraordinary orange cake; therefore, you are interested in the recipe so you can prepare it again. In

principle, it should be an easy task to describe the different steps of cooking and mix all the ingredients. Still, even when you transfer the same instructions is likely, that you will take a few attempts to achieve something similar and as tasty and fluffy. However, this is not as simple as it seems since it can range from a top-level description to a really detailed sequence. At some point in the recipe for, say, *Mediterranean orange cake*, one is instructed to cut four oranges into slices. While to humans, there is no need to provide really detailed instructions (take a knife, sharpen it; clean it up; place the oranges on a wooden board; hold the knife; etc.), for an algorithm, the level of task-related information might be a big difference. Thus, there is a necessary level of granularity when sketching a course of action for a system to follow or a temporal difference model to describe behaviours. Perhaps it is not yet a concern in the models specified along with all the studies of this thesis, but it challenges its findings. This level of detail can be extremely challenging to integrate mathematically speaking to represent more elaborated behavioural patterns (Collins and Frank, 2013).

In fact, evolutionary and comparative psychology demonstrates that primates, children, and adults all rely on similar cognitive mechanisms for learning, which is decomposing complex tasks into a simpler sequence of steps (Spelke and Kinzler, 2007). Indeed, toddlers use the information available to establish sub-goals while playing open-ended games such as building-up block constructions. Toddlers appear to develop sub-goal-guided behaviours in order to accomplish overarching objectives (Kulkarni et al., 2016; Nachum et al., 201 8). But, as happens in the example of the orange cake recipe, the level of information about tasks could easily go from chopping an orange in half to fine slices with a specific kind of knife.

The level of granularity is a central factor in determining the complexity and, therefore, the efficiency of a temporal model in describing behaviours. Once added to RL models, this becomes fundamental since, depending on the knowledge and domain of a task, an agent would need more or less data for a supervised learning setup (Arulkumaran et al., 2017). Finding the most efficient strategy to acquire a goal is a difficult task and can be challenging to specify in computational terms. If granularity can be already demanding in controlled and conventional RL experimental conditions it gets more demanding when you add hierarchies and more ecological settings, as can happen in a simple meal recipe (Momennejad et al., 2017). It remains an open question whether the HRL algorithm used in Studies 2 and 3 would be suitable in the experiments of the present thesis in case of larger sessions or additional hierarchy or task levels. Indeed, the RL approach used in the present thesis is relatively simple compared to sophisticated RL approaches, which have demonstrated outstanding results, such as outperforming humanity's best at Go, learning to play Atari games, and training computers in simulations or in the real world (Silver et al., 2018; Mnih et al., 2015). These accomplishments represent a significant progression in trial-and-error learning but still seem limited in describing human behaviour. While HRL seeks to deconstruct complex issues into smaller ones (efficient learning; Bacon et al., 2017), there is still a way to scale up to ecological circumstances, such as the transferable knowledge of baking an orange cake. The findings of this thesis strive in that direction, tracking down the choice behaviour of human agents trying to maximise rewards in a hierarchical setting.

Therefore, there is a real challenge to generalise the temporal abstraction strategy of an agent toward a specific task. The decomposition in sequences could result in extremely lengthy pathways between the beginning and goal

states. For instance, the recipe to prepare the orange cake could be extremely lengthy or really short; either case would alter the way an algorithm could predict learning, therefore, choice behaviour. In fact, the length of these sub-tasks influences the cost of learning. Therefore, the solution to resolve the dimensionality curse might become another limitation where learning speeds up but not necessarily in an efficient way towards an ultimate reward (Nachum et al., 2018). After revising the persistent findings of studies 1, 2 and 3, one might consider that hierarchical learning is not a guarantee of learning optimal behavioural strategies since it could potentially challenge the conventional principles of optimal choice behaviour and learning. However, research has demonstrated that hierarchical methods such as HRL can significantly reduce the computational costs associated with finding the optimal behavioural strategies (Botvinick, 2012).

Additionally, there is a substantial body of work on sub-goal discovery (Schmidhuber and Wiering, 1996), intrinsic drive (Oudeyer and Kaplan, 2009), and induced curiosity (Fruit and Lazaric, 2017). Nonetheless, there is still a lack of a standard way for incorporating hierarchies into the efficient RL algorithms, given that, depending on the framework being used, manually creating a decent hierarchy needs domain-specific expertise and careful engineering for an algorithm to accurately represent complex behaviours (Collins and Frank, 2013; Eickstein and Collins, 2020).

To choose an acceptable hierarchy structure to represent a task, one must first consider the availability of domain knowledge. Besides this potential mathematical constraint, the other side of the coin might be that HRL is good at describing complex behaviours, and probable consequences of hierarchical abstraction (inefficient learning and sub-optimal choice behaviours) are, to some extent, reflected in the computations. Therefore, an imperfect

algorithm just describes a non-optimal behaviour of agents striving to maximise rewards and relying on, perhaps excessively, auxiliary tasks to achieve a superordinate goal. In this sense, more research needs to be conducted to identify the individual differences between agents in a specific environment (humans of different ages, animals of different species, robots with different purposes, etc.). Therefore, in order to ensure accuracy and an ecological algorithmic architecture, we would attend more tailor-made algorithms rather than standardised criteria to predefined, for instance, learning and decision making (Eckstein and Collins, 2020).

Limitations and Future Directions

The main results of this thesis corroborated the importance of sub-goals and pseudo-feedback processing in Learning and Decision Making. Indeed, the studies have presented the immense potential of pseudo-feedback to speed-up learning and guide behaviours, even in scenarios where the pursuance of sub-goals might lead to a sub-optimal choice behaviour pattern. In the following, we will discuss ideas to embrace future research on this matter.

To start, it is important to remark that in addition to learning how to respond to diverse circumstances and then deciding to pick a course of action, humans learn to arrange their experiences into mental representations that help future behaviours (Radulescu et al., 2021). This representation learning approach encompasses attention and memory as drivers of how an individual explores the world, and frequently, such interaction brings attentional bias (Ma et al., 2018). Indeed, humans depend on multiple types

of statistical inference to organise prior experiences, and this inference process gives birth to compact representations of tasks that guide behaviours in complicated situations. The potential intersection between HRL and the representation learning theories could be addressed as both fields of work trying to understand how individuals break down problems into simple and achievable tasks (Radulescu et al., 2021).

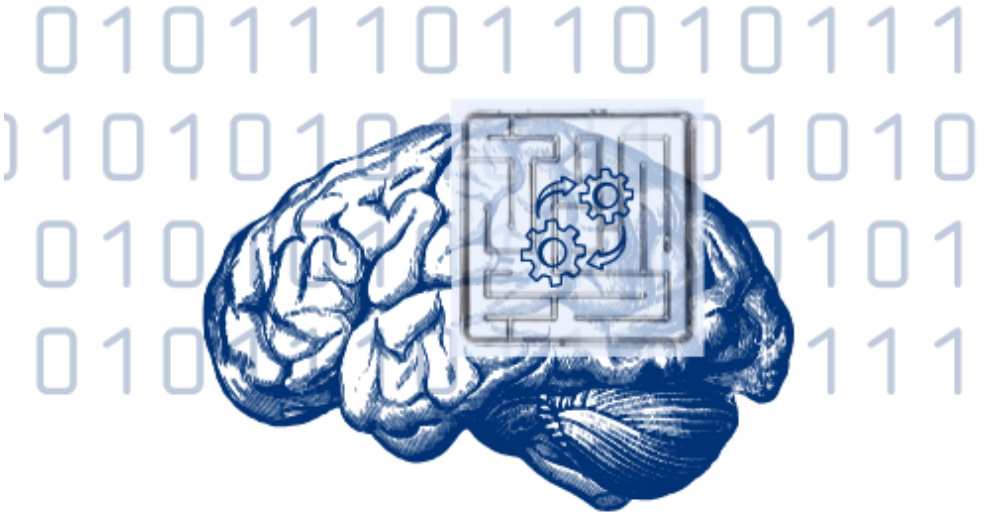
Along with the studies 1, 2 and 3, the data collection was conducted with a very specific population, university students. The bias towards the most-pseudo rewarding options was consistent, even when the tasks were lessening the final rewards for such preference but has not been contrasted with other populations. For instance, research on addictions demonstrates an increased proclivity for "impulsive" behaviour, as well as increased susceptibility to rewards of different kinds (Coffey et al., 2003; Roesch et al., 2007). Indeed, relying on a prediction error that is dominated by immediate rewards might enhance impulsivity, potentially as it happens with pseudo-feedback along with the whole thesis, but also as it has been reported in patients with addiction to cocaine (Ma et al., 2018). According to Hester and Luijten (2014), drug abuse could lead to reducing the typical activity of the Anterior Cingulate Cortex (ACC). There is no concrete evidence that the prefrontal areas are more important in attentional bias than the reward regions such striatum. Therefore, the effect on pseudo-feedback might contribute to explaining the gap between theory and human imaging research involving attentional bias neural networks (Ma et al., 2018). To extend our knowledge about the pseudo-rewards' role in learning and decision making, more comparative research should be conducted on pathologies with deficits in impulsivity such as, for example as drug-addiction, attention deficit hyperactivity disorder, borderline personality disorder, gambling disorders or bipolar disorder, among others. For example, the study of how pathological

gamblers (characterised as having high levels of impulsiveness; Potenza, 2013) encode pseudo reward prediction errors can be an important line of research to extend our knowledge of sub-goal-directed behaviours. Actually, in the clinical literature, the coexistence between anxiety disorders and addictive behaviours is extensively recognised. Psychological theories of addictive and risky behaviours presume the presence of incidental factors such as appetite (impulsivity) and mood control (related to anxiety; Raymond et al., 2003; Gola et al., 2015). Perhaps in the cases of populations with high levels of anxiety and impulsivity, pseudo-feedbacks could generate biases quicker than in a normal population. Also, and departing from the fact that human neuroimaging research has examined dysfunctions of the brain's reward system, specifically the ventral striatum activity underlying anxiety and impulsivity, further experiments- in neurological patients, for instance, patients with damage in the medial prefrontal cortex and/or ventral striatum, could elucidate the role that pseudo-feedback could have in clinical treatments and therapeutic strategies (Pujara et al., 2016).

Another limitation of the present thesis is the proposed dissociation between the roles of ACC and VS. Therefore, in the thesis, it has been proposed that ACC would be in charge of learning action selection while the striatum would play a role in learning, which would be primarily concerned with responding to the value of the immediate reward (Holroyd and Yeung, 2012; Holroyd and Umemoto, 2016). However, it is important to note that both VS and ACC (as it was indexed by theta power) activities reflected both pseudo-reward and reward prediction errors. Therefore, future studies should examine in parallel the ACC and striatum activity to disentangle more clearly the roles of these two systems in HRL. Perhaps simultaneous EEG and fMRI recording could shed light on the common dynamics and

connection between temporal mechanisms such as the registered theta oscillations and the activation of other brain regions.

In the study 2, we used an HRL model to describe the behavioural patterns in terms of sub-optimal choice behaviour. Besides the proposal of adding an additional learning rate to each process running in parallel (encoding of reward prediction errors at different levels), additional research needs to be conducted to establish a more suitable algorithm to mirror ecological hierarchical behaviours during sub-optimal decision-making (Botvinick et al., 2015). One constraint might be that HRL models could suffer from flaws that impair learning and make them unsuitable for application in more realistic day-to-day contexts. HRL tries to alleviate precisely learning difficulty by dividing learning down into more discrete components (Bacon et al., 2017). Indeed, a potential source of algorithmic inaccuracy could be that sub-tasks and abstract actions can be employed in as many tasks as the agent might require based on the history of rewards, former preferences, and individual characteristics, and this embedded variability could suppose a strong mathematical constraint. Therefore, more research should be done on these algorithms to improve such limitations to be applied to approaches to real-life phenomena. In this sense, the fact that no significant individual differences were found to be related to the parameters of the proposed model would support the idea that further refinements are needed to be able to capture the complexity of the studied phenomenon.



Chapter 7. Conclusion

Chapter 7: Conclusion

In this dissertation, we present new evidence regarding the oscillatory mechanisms and brain activity involved in biased decision-making during the resolution of a hierarchical task. In fact, behavioural and computational evidence serves as the basis for describing the choice behaviour of participants from the standpoint of Reinforcement Learning. We obtained behavioural, EEG, and BOLD data regarding how participants pursue sub-goals and encode reward and pseudo-reward prediction errors. Overall, the results demonstrated that pseudo-feedback is capable of inducing prediction errors and influencing choice behaviour, even leading to sub-optimal decision making.

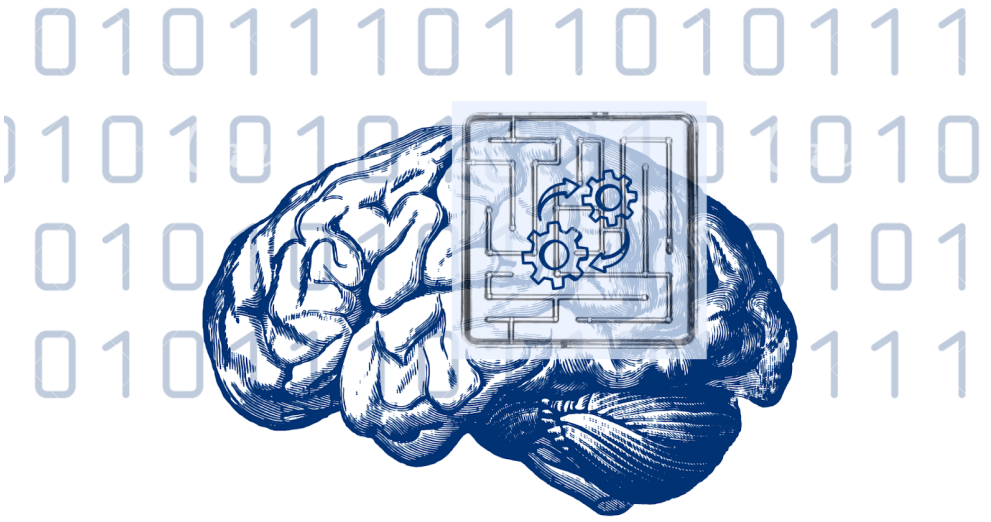
Various conclusions can be outlined after merging the relevant findings from the three studies. Firstly, findings from behavioural and computational modelling indicate that people prefer options with greater pseudo-rewarding properties. Moreover, the findings demonstrated that individuals might make sub-optimal decisions in order to earn more pseudo-rewards without properly monitoring the costs of such preference. Consequently, sub-goals drive learning while providing more accurate predictive values for pre-states that anticipate a final reward to the extent that they generate potential biases and losses.

Also, the oscillatory brain activity in relation to theta power was found to be associated with the pseudo-reward prediction error in study 2. These findings of theta oscillations on the mid-frontal cortex suggest an active role of structures such as the anterior cingulate cortex (ACC) in performance monitoring and action selection, but in this case, not in error detection since

the participants choose a sub-goal directed strategy even when maintaining a sub-optimal pattern of choice.

Finally, study 3 showed parallel processing of both reward and pseudo-reward prediction errors in the Ventral Striatum. This evidence indicates that in a sub-goal-directed behaviour setting where optimality is compromised, the VS plays a crucial role while reinforcing pseudo-feedback attainment.

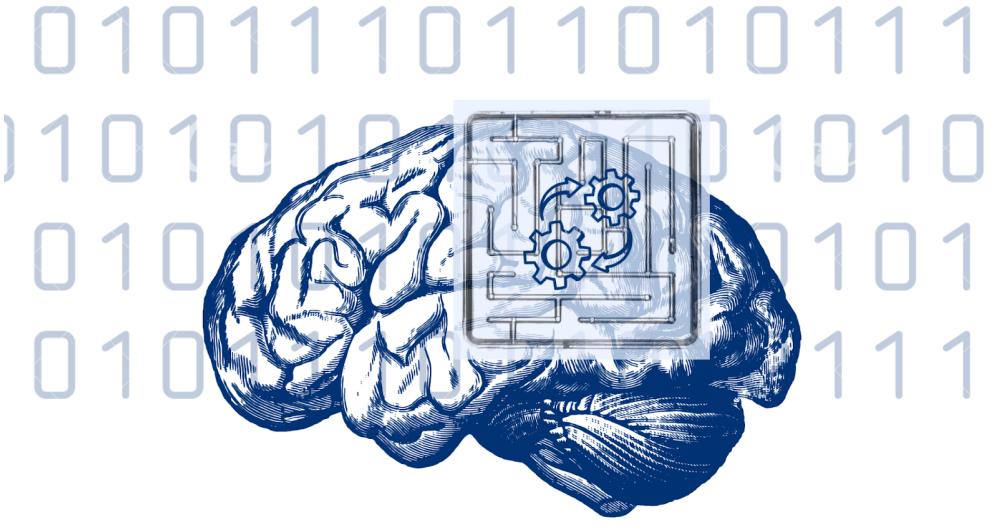
Overall, we demonstrated that sub-goal-directed behaviour, which emerges when complex tasks are broken down into simpler schemes, can lead to inefficient learning and sub-optimal decision-making. In addition, we showed that both VS and ACC were involved in the computation of prediction errors at different hierarchical levels but played a differential role during sub-optimal choice behaviour. The integration of results from electrophysiological, BOLD responses, and temporal difference modelling helps to compose a wider picture of the neural mechanisms involved in pseudo-reward processing during sub-optimal decision-making in complex behaviours.



8. Abbreviation List

8. Abbreviation List.

ACC – Anterior Cingulate Cortex
Amyg – Amygdala
ANOVA – Analysis of Variance
BG – Basal Ganglia
BOLD – Blood-oxygen-level dependent
DA – Dopamine
dlPFC – Dorsolateral Prefrontal Cortex
dpMFC – Dorsal Posterior Medial Frontal Cortex
EEG – Electroencephalography
ERN – Error-related Negativity
ERP – Event-related Potentials
fMRI – Functional Magnetic Resonance
FRN – Feedback-related Negativity
HDI – High Density Interval
HRL – Hierarchical Reinforcement Learning
Hyp – Hypothalamus
MDP – Markov Decision Process
MEG – Magnetoencephalography
OFC – Orbitofrontal Cortex
PE – Prediction Error
PFC – Prefrontal Cortex
pmPFC – Posterior Medial Prefrontal Cortex
PRPE – Pseudo Reward Prediction Error
RL – Reinforcement Learning
ROI – Region of interest
RPE – Reward Prediction Error
SD – Standard Deviation
SE – Standard Error
TD – Temporal Difference
vmPFC – Ventromedial Prefrontal Cortex
VS – Ventral Striatum
VTA – Ventral Tegmental Area



9. References

9. References.

Adkins, T. J., and Lee, T. G. (2021). Reward modulates cortical representations of action. Neuroimage, 228, 117708.

Amaya, K. A., Stott, J. J., and Smith, K. S. (2020). Sign-tracking behaviour is sensitive to outcome devaluation in a devaluation context-dependent manner: implications for analyzing habitual behaviour. Learning and Memory, 27(4), 136-149.

Andreou, C., Frielinghaus, H., Rauh, J., Mußmann, M., Vauth, S., Braun, P., and Mulert, C. (2017). Theta and high-beta networks for feedback processing: a simultaneous EEG–fMRI study in healthy male subjects. *Translational Psychiatry*, *7*(1), e1016-e1016.

Apitz, T., and Bunzeck, N. (2014). Early effects of reward anticipation are modulated by dopaminergic stimulation. *PloS one*, *9*(10), e108886.

Arulkumaran, K., Deisenroth, M. P., Brundage, M., and Bharath, A. A. (2017). A brief survey of deep reinforcement learning. *arXiv preprint arXiv:1708.05866*.

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

Bacon, P. L., Henderson, P., Chang, W. D., Meger, D., Pineau, J., and Precup, D. (2017). Optiongan: Learning joint reward-policy options using generative adversarial inverse reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).

Badre, D. (2008). Cognitive control, hierarchy, and the rostro–caudal organization of the frontal lobes. *Trends in cognitive sciences*, *12*(5), 193-200.

Balaguer, J., Spiers, H., Hassabis, D., and Summerfield, C. (2016). Neural mechanisms of hierarchical planning in a virtual subway network. *Neuron*, *90*(4), 893-903.

Bar-Gad, I., Havazelet-Heimer, G., Goldberg, J. A., Ruppin, E., and Bergman, H. (2000). Reinforcement-driven dimensionality reduction-a model for information processing in the basal ganglia. *Journal of basic and clinical physiology and pharmacology*, *11*(4), 305-320.

Bates D, Mächler M, Bolker B, Walker S (2015). "Fitting Linear Mixed-Effects Models Using lme4." Journal of Statistical Software, 67(1), 1–48, 10.18637/jss.v067.i01.

Bechara, A., and Damasio, H. (2002). Decision-making and addiction (part I): impaired activation of somatic states in substance dependent individuals when pondering decisions with negative future consequences. *Neuropsychologia*, *40*(10), 1675-1689.

Behrens, T. E., and Jocham, G. (2011). How to perfect a chocolate souffle and other important problems. *Neuron*, *71*(2), 203-205.

Beierholm, U. R., and Dayan, P. (2010). Pavlovian-instrumental interaction in 'observing behaviour'. *PLoS computational biology*, *6*(9), e1000903.

Beierholm, U. R., Wunderlich, K., Bossaerts, P., and O'Doherty, J. P. (2011). The human prefrontal cortex mediates integration of potential causes behind observed outcomes. *Journal of neurophysiology*, *106*(3), 1558-1569.

Bernat, E. M., Nelson, L. D., and Baskin-Sommers, A. R. (2015). Time-frequency theta and delta measures index separable components of feedback processing in a gambling task. *Psychophysiology*, *52*(5), 626-637.

Berridge, K. C. (2007). The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology*, *191*(3), 391-431.

Bogacz, R. (2007). Optimal decision-making theories: linking neurobiology with behaviour. *Trends in cognitive sciences*, *11*(3), 118-125.

Botvinick, M. M. (2012). Hierarchical reinforcement learning and decision making. *Current opinion in neurobiology*, *22*(6), 956-962.

Botvinick, M. M. , , Solway, A., Diuk, C., McGuire, J. T., Barto, A. G., Niv, Y., and Ribas-Fernandes, J. J. (2011). A neural signature of hierarchical reinforcement learning. *Neuron*, *71*(2), 370-379.

Botvinick, M. M., Huffstetler, S., and McGuire, J. T. (2009). Effort discounting in human nucleus accumbens. *Cognitive, affective, and behavioural neuroscience*, *9*(1), 16-27.

Botvinick, M. M., Niv, Y., and Barto, A. G. (2009). Hierarchically organized behaviour and its neural foundations: A reinforcement learning perspective. *Cognition*, *113*(3), 262-280.

Botvinick, M., and Weinstein, A. (2014). Model-based hierarchical reinforcement learning and human action control. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), 20130480.

Botvinick, M., Weinstein, A., Solway, A., and Barto, A. (2015). Reinforcement learning, efficient coding, and the statistics of natural tasks. *Current opinion in behavioural sciences*, *5*, 71-77.

Brandtstädter, J., and Rothermund, K. (2002). The life-course dynamics of goal pursuit and goal adjustment: A two-process framework. *Developmental review*, *22*(1), 117-150.

Broche-Pérez, Y., Jiménez, L. H., and Omar-Martínez, E. (2016). Neural substrates of decision-making. *Neurología (English Edition)*, *31*(5), 319-325.

Brogan, D. J., Brown, G. G., Williams, R. L., and Bieler, G. S. (2010). Estimating model-adjusted risks, risk differences, and risk ratios from complex survey data. *American journal of epidemiology*, *171*(5), 618-623.

Bromberg-Martin, E. S., and Hikosaka, O. (2009). Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron*, *63*(1), 119-126.

Cai, W., and Leung, H. C. (2011). Rule-guided executive control of response inhibition: functional topography of the inferior frontal cortex. *PloS one*, *6*(6), e20840.

Caligiore, D., Arbib, M. A., Miall, R. C., and Baldassarre, G. (2019). The super-learning hypothesis: Integrating learning processes across cortex, cerebellum and basal ganglia. Neuroscience and Biobehavioural Reviews, 100, 19-34.

Canas, F., and Jones, M. (2010). Attention and reinforcement learning: constructing representations from indirect feedback. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 32, No. 32).

Castegnetti, G., Zurita, M., and De Martino, B. (2021). How usefulness shapes neural representations during goal-directed behaviour. Science Advances, 7(15), eabd5363.

Cavanagh, J. F., and Frank, M. J. (2014). Frontal theta as a mechanism for cognitive control. *Trends in cognitive sciences*, *18*(8), 414-421.

Cavanagh, J. F., Frank, M. J., Klein, T. J., and Allen, J. J. (2010). Frontal theta links prediction errors to behavioural adaptation in reinforcement learning. *Neuroimage*, *49*(4), 3198-3209.

Cavanagh, J. F., Wiecki, T. V., Cohen, M. X., Figueroa, C. M., Samanta, J., Sherman, S. J., and Frank, M. J. (2011). Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nature neuroscience*, *14*(11), 1462-1467.

Chiang, F. K., and Wallis, J. D. (2018). Neuronal encoding in prefrontal cortex during hierarchical reinforcement learning. *Journal of cognitive neuroscience*, *30*(8), 1197-1208.

Chow, J. J., Smith, A. P., Wilson, A. G., Zentall, T. R., and Beckmann, J. S. (2017). Sub-optimal choice in rats: Incentive salience attribution promotes maladaptive decision-making. *Behavioural Brain Research*, *320*, 244-254.

Clark, R., and Gilchrist, I. D. (2018). The relationship between reward and probability: evidence that exploration may be intrinsically rewarding. *Visual Cognition*, *26*(9), 672-694.

Coffey, S. F., Gudleski, G. D., Saladin, M. E., and Brady, K. T. (2003). Impulsivity and rapid discounting of delayed hypothetical rewards in cocaine-dependent individuals. *Experimental and clinical psychopharmacology*, *11*(1), 18.

Cohen, M. X. (2017). Comparison of linear spatial filters for identifying oscillatory activity in multichannel data. *Journal of neuroscience methods*, *278*, 1-12.

Cohen, M. X., and Cavanagh, J. F. (2011). Single-trial regression elucidates the role of prefrontal theta oscillations in response conflict. *Frontiers in psychology*, *2*, 30.

Cohen, M. X., and Frank, M. J. (2009). Neurocomputational models of basal ganglia function in learning, memory and choice. *Behavioural brain research*, *199*(1), 141-156.

Colaizzi, J. M., Flagel, S. B., Joyner, M. A., Gearhardt, A. N., Stewart, J. L., and Paulus, M. P. (2020). Mapping sign-tracking and goal-tracking onto human behaviours. *Neuroscience and Biobehavioural Reviews*, *111*, 84-94.

Collins, A. G. (2018). The tortoise and the hare: Interactions between reinforcement learning and working memory. *Journal of cognitive neuroscience*, *30*(10), 1422-1432.

Collins, A. G., and Cockburn, J. (2020). Beyond dichotomies in reinforcement learning. *Nature Reviews Neuroscience*, *21*(10), 576-586.

Collins, A. G., and Frank, M. J. (2013). Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychological review*, *120*(1), 190.

Collins, A. G., and Frank, M. J. (2014). A reinforcement learning mechanism responsible for the valuation of free choice. *Neuron*, *83*(3), 551-557.

Collins, A. G., Ciullo, B., Frank, M. J., and Badre, D. (2017). Working memory load strengthens reward prediction errors. *Journal of Neuroscience*, *37*(16), 4332-4342.

Corbetta, M., and Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, *3*(3), 201-215.

Cunningham, P. J. (2020). *A Temporal Information-Theoretic Model of Sub-optimal Choice* (Doctoral dissertation, Utah State University).

d'Acremont, M., and Bossaerts, P. (2008). Neurobiological studies of risk assessment: a comparison of expected utility and mean-variance approaches. Cognitive, Affective, and Behavioural Neuroscience, 8(4), 363-374.

D'Ardenne, K., Hennigan, K., and McClure, S. M. (2015). Distinct midbrain and habenula pathways are involved in processing aversive events in humans. *Journal of Neuroscience*, *35*(1), 198-208.

Daw, N. D. (2015). Of goals and habits. Proceedings of the National Academy of Sciences, 112(45), 13749-13750.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., and Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204-1215.

Daw, N. D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioural control. *Nature neuroscience*, *8*(12), 1704-1711.

Dayan, P., and Balleine, B. W. (2002). Reward, motivation, and reinforcement learning. *Neuron*, *36*(2), 285-298.

Dayan, P., and Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. *Current opinion in neurobiology*, *18*(2), 185-196.

De Lange, F. P., Heilbron, M., and Kok, P. (2018). How do expectations shape perception?. Trends in cognitive sciences, 22(9), 764-779.

Decker, J. H., Otto, A. R., Daw, N. D., and Hartley, C. A. (2016). From creatures of habit to goal-directed learners: Tracking the developmental emergence of model-based reinforcement learning. *Psychological science*, *27*(6), 848-858.

Decker, J. H., Otto, A. R., Daw, N. D., and Hartley, C. A. (2016). From creatures of habit to goal-directed learners: Tracking the developmental emergence of model-based reinforcement learning. *Psychological science*, *27*(6), 848-858.

Delgado, M. R. (2007). Reward-related responses in the human striatum. Annals of the New York Academy of Sciences, 1104(1), 70-88.

Den Ouden, H. E., Kok, P., and De Lange, F. P. (2012). How prediction errors shape perception, attention, and motivation. *Frontiers in psychology*, *3*, 548.

Dezfouli, A., and Balleine, B. W. (2012). Habits, action sequences and reinforcement learning. *European Journal of Neuroscience*, *35*(7), 1036-1051.

Dickinson, A., Wood, N., and Smith, J. W. (2002). Alcohol seeking by rats: action or habit?. *The Quarterly Journal of Experimental Psychology Section B*, *55*(4b), 331-348.

Diekhof, E. K., Kaps, L., Falkai, P., and Gruber, O. (2012). The role of the human ventral striatum and the medial orbitofrontal cortex in the representation of reward magnitude–An activation likelihood estimation meta-analysis of neuroimaging studies of passive reward expectancy and outcome processing. *Neuropsychologia*, *50*(7), 1252-1266.

Diuk, C., Schapiro, A., Córdova, N., Ribas-Fernandes, J., Niv, Y., and Botvinick, M. (2013a). Divide and conquer: hierarchical reinforcement learning and task

decomposition in humans. In *Computational and robotic models of the hierarchical organization of behaviour* (pp. 271-291). Springer, Berlin, Heidelberg.

Diuk, C., Tsai, K., Wallis, J., Botvinick, M., and Niv, Y. (2013b). Hierarchical learning induces two simultaneous, but separable, prediction errors in human basal ganglia. *Journal of Neuroscience*, *33*(13), 5797-5805.

Dorfman, H. M., and Gershman, S. J. (2019). Controllability governs the balance between Pavlovian and instrumental action selection. *Nature communications*, *10*(1), 1-8.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex?. *Neural networks*, *12*(7-8), 961-974.

Doya, K., Miyazaki, K. W., and Miyazaki, K. (2012). Activation of dorsal raphe serotonin neurons is necessary for waiting for delayed rewards. *Journal of Neuroscience*, *32*(31), 10451-10457.

Dulac-Arnold, G., Mankowitz, D., and Hester, T. (2019). Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*.

Eckstein, M. K., and Collins, A. G. (2020). Computational evidence for hierarchically structured reinforcement learning in humans. *Proceedings of the National Academy of Sciences*, *117*(47), 29381-29389.

Eliaz, K., and Schotter, A. (2010). Paying for confidence: An experimental study of the demand for non-instrumental information. *Games and Economic Behaviour*, *70*(2), 304-324.

Erdeniz, B., Garrison, J., and Done, J. (2013). Prediction error in reinforcement learning: a meta-analysis of neuroimaging studies. *Neuroscience and Biobehavioural Reviews*, *37*(7), 1297-1310.

FitzGerald, T. H., Friston, K. J., and Dolan, R. J. (2012). Action-specific value signals in reward-related regions of the human brain. Journal of Neuroscience, 32(46), 16417-16423.

Foti, D., Carlson, J. M., Harmon-Jones, E., and Proudfit, G. H. (2015). Midbrain volume predicts fMRI and ERP measures of reward reactivity. *Brain Structure and Function*, *220*(3), 1861-1866.

Foti, D., Masaki, H., Hirao, T., Maruo, Y., and Hajcak, G. (2018). Feedback-related electroencephalogram oscillations of athletes with high and low sports anxiety. *Frontiers in psychology*, *9*, 1420.

Frank, M. J., and Badre, D. (2012). Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: computational analysis. *Cerebral cortex*, *22*(3), 509-526.

Fruit, R., and Lazaric, A. (2017, April). Exploration-exploitation in mdps with options. In *Artificial intelligence and statistics* (pp. 576-584). PMLR.

Fu, W. T., and Gray, W. D. (2006). Sub-optimal tradeoffs in information seeking. *Cognitive Psychology*, *52*(3), 195-242.

Gardner, M. P., Conroy, J. C., Sanchez, D. C., Zhou, J., and Schoenbaum, G. (2019). Real-time value integration during economic choice is regulated by orbitofrontal cortex. *Current Biology*, *29*(24), 4315-4322.

Garofalo, G., Pellegrino, E., Lorini, C., Allodi, G., Buonamici, C., and Bonaccorsi, G. (2013). Music-listening habits with MP3 player in a group of adolescents: a descriptive survey. *Annali di igiene: medicina preventiva e di comunita*, *25*(5), 367-376.

Gebhardt, C., Oulasvirta, A., and Hilliges, O. (2021). Hierarchical reinforcement learning explains task interleaving behaviour. *Computational Brain and Behaviour*, *4*(3), 284-304.

Gelman, A., and Rubin, D. B. (1992). A single series from the Gibbs sampler provides a false sense of security. *Bayesian statistics*, *4*, 625-631.

Gheza, D., Bakic, J., Baeken, C., De Raedt, R., and Pourtois, G. (2019). Abnormal approach-related motivation but spared reinforcement learning in MDD:

Evidence from fronto-midline Theta oscillations and frontal Alpha asymmetry. Cognitive, Affective, and Behavioural Neuroscience, 19(3), 759-777.

Gheza, D., De Raedt, R., Baeken, C., and Pourtois, G. (2018). Integration of reward with cost anticipation during performance monitoring revealed by ERPs and EEG spectral perturbations. *NeuroImage*, *173*, 153-164.

Glazer, J. E., Kelley, N. J., Pornpattananangkul, N., Mittal, V. A., and Nusslock, R. (2018). Beyond the FRN: Broadening the time-course of EEG and ERP components implicated in reward processing. International Journal of Psychophysiology, 132, 184-202.

Glimcher, P. W., and Fehr, E. (Eds.). (2013). Neuroeconomics: Decision making and the brain. Academic Press.

Glöckner, A., and Pachur, T. (2012). Cognitive models of risky choice: Parameter stability and predictive accuracy of prospect theory. Cognition, 123(1), 21-32.

Gola, M., Miyakoshi, M., and Sescousse, G. (2015). Sex, impulsivity, and anxiety: Interplay between ventral striatum and amygdala reactivity in sexual behaviours. Journal of Neuroscience, 35(46), 15227-15229.

Gold, J. I., and Shadlen, M. N. (2007). The neural basis of decision making. Annu. Rev. Neurosci., 30, 535-574.

Graybiel, A. M., and Grafton, S. T. (2015). The striatum: where skills and habits meet. *Cold Spring Harbor perspectives in biology*, *7*(8), a021691.

Gremel, C. M., and Costa, R. M. (2013). Orbitofrontal and striatal circuits dynamically encode the shift between goal-directed and habitual actions. *Nature communications*, *4*(1), 1-12.

Gruber, M. J., Gelman, B. D., and Ranganath, C. (2014). States of curiosity modulate hippocampus-dependent learning via the dopaminergic circuit. Neuron, 84(2), 486-496.

Haber, S. N. (2017). Anatomy and connectivity of the reward circuit. In Decision neuroscience (pp. 3-19). Academic Press.

Haber, S. N., and Knutson, B. (2010). The reward circuit: linking primate anatomy and human imaging. *Neuropsychopharmacology*, *35*(1), 4-26.

HajiHosseini, A., Rodríguez-Fornells, A., and Marco-Pallarés, J. (2012). The role of beta-gamma oscillations in unexpected rewards processing. Neuroimage, 60(3), 1678-1685.

Hammers, A., Allom, R., Koepp, M. J., Free, S. L., Myers, R., Lemieux, L., and Duncan, J. S. (2003). Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. Human brain mapping, 19(4), 224-247.

Hassall, C. D. (2019). The neural correlates of exploration (Doctoral dissertation).

Hayden, B. Y., and Niv, Y. (2021). The case against economic values in the orbitofrontal cortex (or anywhere else in the brain). Behavioural Neuroscience, 135(2), 192.

Heekeren, H. R., Wartenburger, I., Marschner, A., Mell, T., Villringer, A., and Reischies, F. M. (2007). Role of ventral striatum in reward-based decision making. Neuroreport, 18(10), 951-955.

Hengst, B. (2012). Hierarchical approaches. In Reinforcement learning (pp. 293-323). Springer, Berlin, Heidelberg.

Holroyd, C. B., and Coles, M. G. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. Psychological review, 109(4), 679.

Holroyd, C. B., and Umemoto, A. (2016). The research domain criteria framework: The case for anterior cingulate cortex. Neuroscience and Biobehavioural Reviews, 71, 418-443.

Holroyd, C. B., and Yeung, N. (2012). Motivation of extended behaviours by anterior cingulate cortex. Trends in cognitive sciences, 16(2), 122-128.

Holroyd, C. B., Hyman, J. M., and Seamans, J. K. (2017). A novel neural prediction error found in anterior cingulate cortex ensembles. *Neuron*, *95*(2), 447-456.

Houk, J. C., and Wise, S. P. (1995). Distributed modular architectures linking basal ganglia, cerebellum, and cerebral cortex: their role in planning and controlling action. *Cerebral cortex*, *5*(2), 95-110.

Huang, Y., Yaple, Z. A., and Yu, R. (2020). Goal-oriented and habitual decisions: Neural signatures of model-based and model-free learning. *NeuroImage*, *215*, 116834.

Iglesias, S., Mathys, C., Brodersen, K. H., Kasper, L., Piccirelli, M., den Ouden, H. E., and Stephan, K. E. (2013). Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. Neuron, 80(2), 519-530.

Jenison, R. L., Rangel, A., Oya, H., Kawasaki, H., and Howard, M. A. (2011). Value encoding in single neurons in the human amygdala during decision making. Journal of Neuroscience, 31(1), 331-338.

Kahneman, D., and Tversky, A. (1979). On the interpretation of intuitive probability: A reply to Jonathan Cohen.

Kahneman, D., and Tversky, A. (1983). Can irrationality be intelligently discussed. Behavioural and Brain Sciences, 6(3), 509-510.

Kahnt, T., and Tobler, P. N. (2017). Reward, value, and salience. In *Decision neuroscience* (pp. 109-120). Academic Press.

Kamigaki, T. (2019). Prefrontal circuit organization for executive control. *Neuroscience research*, *140*, 23-36.

Kim, H., Sul, J. H., Huh, N., Lee, D., and Jung, M. W. (2009). Role of striatum in updating values of chosen actions. Journal of neuroscience, 29(47), 14701-14712.

Kirby, K. N., Petry, N. M., and Bickel, W. K. (1999). Heroin addicts have higher discount rates for delayed rewards than non-drug-using controls. Journal of Experimental psychology: general, 128(1), 78.

Klein-Flügge, M. C., Hunt, L. T., Bach, D. R., Dolan, R. J., and Behrens, T. E. (2011). Dissociable reward and timing signals in human midbrain and ventral striatum. Neuron, 72(4), 654-664.

Kolling, N., Behrens, T. E., Wittmann, M. K., and Rushworth, M. F. (2016). Multiple signals in anterior cingulate cortex. Current opinion in neurobiology, 37, 36-43.

Konidaris, G., and Barto, A. (2009). Skill discovery in continuous reinforcement learning domains using skill chaining. Advances in neural information processing systems, 22.

Kreps, D. M., and Porteus, E. L. (1978). Temporal resolution of uncertainty and dynamic choice theory. Econometrica: journal of the Econometric Society, 185-200.

Kruschke, J. (2014). Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan.

Kulkarni, T. D., Narasimhan, K., Saeedi, A., and Tenenbaum, J. (2016). Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. Advances in neural information processing systems, 29.

Lak, A., Okun, M., Moss, M. M., Gurnani, H., Farrell, K., Wells, M. J., and Carandini, M. (2020). Dopaminergic and prefrontal basis of learning from sensory confidence and reward value. Neuron, 105(4), 700-711.

Lee, D., Seo, H., and Jung, M. W. (2012). Neural basis of reinforcement learning and decision making. Annual review of neuroscience, 35, 287-308.

Lee, S. W., Shimojo, S., and O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. Neuron, 81(3), 687-699.

Lega, B. C., Kahana, M., Jaggi, J., Baltuch, G. H., and Zaghloul, K. (2011). Neuronal and oscillatory activity during reward processing in the human ventral striatum. Neuroreport, 22(16), 795.

Lesaint, F., Sigaud, O., Clark, J. J., Flagel, S. B., and Khamassi, M. (2015). Experimental predictions drawn from a computational model of sign-trackers and goal-trackers. Journal of Physiology-Paris, 109(1-3), 78-86.

Levine, S. (2018). Reinforcement learning and control as probabilistic inference: Tutorial and review. arXiv preprint arXiv:1805.00909.

Levy, D. J., and Glimcher, P. W. (2011). Comparing apples and oranges: using reward-specific and reward-general subjective value representation in the brain. Journal of Neuroscience, 31(41), 14693-14707.

Liakoni, V., Lehmann, M. P., Modirshanechi, A., Brea, J., Lutti, A., Gerstner, W., and Preuschoff, K. (2022). Brain signals of a Surprise-Actor-Critic model: Evidence for multiple learning modules in human decision making. NeuroImage, 246, 118780.

Louie, K., Holper, L., Van Brussel, L. D., Schmidt, L., Schulthess, S., Burke, C. J., and Tobler, P. N. (2017). Adaptive value normalization in the prefrontal cortex is reduced by memory load. eneuro, 4(2).

Lüscher, C., and Janak, P. H. (2021). Consolidating the circuit model for addiction. *Annual Review of Neuroscience*, *44*, 173-195.

Ma, L., Steinberg, J. L., Cunningham, K. A., Bjork, J. M., Lane, S. D., Schmitz, J. M., and Moeller, F. G. (2018). Altered anterior cingulate cortex to hippocampus effective connectivity in response to drug cues in men with cocaine use disorder. Psychiatry Research: Neuroimaging, 271, 59-66.

Ma, S. Y., Wegiel, J., Flory, M., Kuchna, I., Nowicki, K., Imaki, H., and Brown, W. T. (2014). Stereological study of the neuronal number and volume of 38 brain subdivisions of subjects diagnosed with autism reveals significant alterations restricted to the striatum, amygdala and cerebellum. Acta neuropathologica communications, 2(1), 1-18.

Małecka, M. (2020). The normative decision theory in economics: A philosophy of science perspective. The case of the expected utility theory. Journal of Economic Methodology, 27(1), 36-50.

Malvaez, M., Shieh, C., Murphy, M. D., Greenfield, V. Y., and Wassum, K. M. (2019). Distinct cortical–amygdala projections drive reward value encoding and retrieval. Nature neuroscience, 22(5), 762-769.

Marche, K., Martel, A. C., and Apicella, P. (2017). Differences between dorsal and ventral striatum in the sensitivity of tonically active neurons to rewarding events. Frontiers in Systems Neuroscience, 11, 52.

Marco-Pallares, J., Cucurell, D., Cunillera, T., García, R., Andrés-Pueyo, A., Münte, T. F., and Rodríguez-Fornells, A. (2008). Human oscillatory activity associated to reward processing in a gambling task. Neuropsychologia, 46(1), 241-248.

Marco-Pallarés, J., Cucurell, D., Cunillera, T., Krämer, U. M., Càmara, E., Nager, W., and Rodriguez-Fornells, A. (2009). Genetic variability in the dopamine system (dopamine receptor D4, catechol-O-methyltransferase) modulates neurophysiological responses to gains and losses. Biological psychiatry, 66(2), 154-161.

Marco-Pallarés, J., Münte, T. F., and Rodríguez-Fornells, A. (2015). The role of high-frequency oscillatory activity in reward processing and learning. Neuroscience and Biobehavioural Reviews, 49, 1-7.

Mas-Herrero, E., and Marco-Pallarés, J. (2014). Frontal theta oscillatory activity is a common mechanism for the computation of unexpected outcomes and learning rate. Journal of cognitive neuroscience, 26(3), 447-458.

Mas-Herrero, E., and Marco-Pallarés, J. (2016). Theta oscillations integrate functionally segregated sub-regions of the medial prefrontal cortex. NeuroImage, 143, 166-174.

Mas-Herrero, E., Sescousse, G., Cools, R., and Marco-Pallares, J. (2019). The contribution of striatal pseudo-reward prediction errors to value-based decision-making. NeuroImage, 193, 67-74.

Matsumoto, T., Fujimoto, T., Takeuch, K., Kamimura, K., Hamada, R., Nakamura, K., and Kato, N. (2007). Abnormal glucose metabolism in the anterior cingulate cortex in patients with schizophrenia. Psychiatry Research: Neuroimaging, 154(1), 49-58.

McDevitt, M. A., Diller, J. W., and Pietrzykowski, M. O. (2019). Human and pigeon sub-optimal choice. Learning and Behaviour, 47(4), 334-343.

McDevitt, M. A., Dunn, R. M., Spetch, M. L., and Ludvig, E. A. (2016). When good news leads to bad choices. Journal of the experimental analysis of behaviour, 105(1), 23-40.

McGovern, A., and Barto, A. G. (2001). Automatic discovery of sub-goals in reinforcement learning using diverse density.

Merel, J., Botvinick, M., and Wayne, G. (2019). Hierarchical motor control in mammals and machines. Nature communications, 10(1), 1-12.

Miletić, S., Boag, R. J., and Forstmann, B. U. (2020). Mutual benefits: Combining reinforcement learning with sequential sampling models. Neuropsychologia, 136, 107261.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., and Hassabis, D. (2015). Human-level control through deep reinforcement learning. nature, 518(7540), 529-533.

Molet, M., Miller, H. C., Laude, J. R., Kirk, C., Manning, B., and Zentall, T. R. (2012). Decision making by humans in a behavioural task: Do humans, like pigeons, show sub-optimal choice?. Learning and behaviour, 40(4), 439-447.

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., and Gershman, S. J. (2017). The successor representation in human reinforcement learning. Nature human behaviour, 1(9), 680-692.

Montague, P. R., King-Casas, B., and Cohen, J. D. (2006). Imaging valuation models in human choice. Annu. Rev. Neurosci., 29, 417-448.

Monteiro, T., Vasconcelos, M., and Kacelnik, A. (2015). Irrational choice and the value of information. Scientific reports, 5(1), 1-12.

Morelli, M., Casagrande, M., and Forte, G. (2021). Decision Making: a Theoretical Review. Integrative Psychological and Behavioural Science, 1-21.

Mormann, M., and Russo, J. E. (2021). Does attention increase the value of choice alternatives?. Trends in cognitive sciences, 25(4), 305-315.

Morris, R. W., Dezfouli, A., Griffiths, K. R., and Balleine, B. W. (2014). Action-value comparisons in the dorsolateral prefrontal cortex control choice between goal-directed actions. Nature communications, 5(1), 1-10.

Nachum, O., Gu, S. S., Lee, H., and Levine, S. (2018). Data-efficient hierarchical reinforcement learning. Advances in neural information processing systems, 31.

Neseliler, S. (2018). The neurobehavioural correlates of appetite control and obesity. McGill University (Canada).

Niv, Y. (2009). Reinforcement learning in the brain. Journal of Mathematical Psychology, 53(3), 139-154.

Niv, Y. (2019). Learning task-state representations. Nature neuroscience, 22(10), 1544-1553.

Niv, Y., Joel, D., and Ruppin, E. (2002). Actor–critic models of the basal ganglia: New anatomical and computational perspectives. Neural networks, 15(4-6), 535-547.

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., and Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. Journal of Neuroscience, 35(21), 8145-8157.

Noppeney, U. (2021). Perceptual inference, learning, and attention in a multisensory world. Annual Review of Neuroscience, 44, 449-473.

Noppeney, U. (2021). Perceptual inference, learning, and attention in a multisensory world. *Annual Review of Neuroscience*, *44*, 449-473.

O'Doherty, J. P., E., Dunne, S., Bossaerts, P., and Payzan-LeNestour (2013). The neural representation of unexpected uncertainty during value-based decision making. Neuron, 79(1), 191-201.

O'Doherty, J. P., Cockburn, J., and Pauli, W. M. (2017). Learning, reward, and decision making. Annual review of psychology, 68, 73-100.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science, 304(5669), 452-454.

Osband, I., Doron, Y., Hessel, M., Aslanides, J., Sezener, E., Saraiva, A., and Van Hasselt, H. (2019). Behaviour suite for reinforcement learning. arXiv preprint arXiv:1908.03568.

Otterlo, M. V., and Wiering, M. (2012). Reinforcement learning and markov decision processes. In Reinforcement learning (pp. 3-42). Springer, Berlin, Heidelberg.

Oudeyer, P. Y., and Kaplan, F. (2009). What is intrinsic motivation? A typology of computational approaches. Frontiers in neurorobotics, 1, 6.

Padoa-Schioppa, C., Conen, K. E., Cai, X., and Rustichini, A. (2017). Optimal coding and neuronal adaptation in economic decisions. Nature communications, 8(1), 1-14.

Palminteri, S., Khamassi, M., Joffily, M., and Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning. Nature communications, 6(1), 1-14.

Pastor-Bernier, A., and Cisek, P. (2011). Making choices between rules or between actions. Neuron, 70(3), 382-384.

Pateria, S., Subagdja, B., Tan, A. H., and Quek, C. (2021). Hierarchical reinforcement learning: A comprehensive survey. ACM Computing Surveys (CSUR), 54(5), 1-35.

Paul, K., Vassena, E., Severo, M. C., and Pourtois, G. (2020). Dissociable effects of reward magnitude on fronto-medial theta and FRN during performance monitoring. Psychophysiology, 57(2), e13481.

Penny, W. D., Stephan, K. E., Daunizeau, J., Rosa, M. J., Friston, K. J., Schofield, T. M., and Leff, A. P. (2010). Comparing families of dynamic causal models. PLoS computational biology, 6(3), e1000709.

Peters, A. J., Fabre, J. M., Steinmetz, N. A., Harris, K. D., and Carandini, M. (2021). Striatal activity topographically reflects cortical activity. Nature, 591(7850), 420-425.

Pezzulo, G., Rigoli, F., and Friston, K. J. (2018). Hierarchical active inference: a theory of motivated control. *Trends in cognitive sciences*, *22*(4), 294-306.

Platt, M. L., and Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. Nature, 400(6741), 233-238.

Pleskac, T. J. (2008). Decision making and learning while taking sequential risks. Journal of Experimental Psychology: Learning, Memory, and Cognition, 34(1), 167.

Polanía, R., Krajbich, I., Grueschow, M., and Ruff, C. C. (2014). Neural oscillations and synchronization differentially support evidence accumulation in perceptual and value-based decision making. Neuron, 82(3), 709-720.

Pujara, M. S., Philippi, C. L., Motzkin, J. C., Baskaya, M. K., and Koenigs, M. (2016). Ventromedial prefrontal cortex damage is associated with decreased ventral striatum volume and response to reward. Journal of Neuroscience, 36(18), 5047-5054.

Radulescu, A., Shin, Y. S., and Niv, Y. (2021). Human representation learning. Annual Review of Neuroscience, 44, 253-273.

Raymond, N. C., Coleman, E., and Miner, M. H. (2003). Psychiatric comorbidity and compulsive/impulsive traits in compulsive sexual behaviour. Comprehensive psychiatry, 44(5), 370-380.

Ribas-Fernandes, J. J. R., Shahnazian, D., Holroyd, C. B., and Botvinick, M. M. (2018). Sub-goal-and goal-related prediction errors in medial prefrontal cortex. bioRxiv, 245829.

Ribas-Fernandes, J. J., Shahnazian, D., Holroyd, C. B., and Botvinick, M. M. (2019). Sub-goal-and goal-related reward prediction errors in medial prefrontal cortex. *Journal of cognitive neuroscience*, *31*(1), 8-23.

Ribas-Fernandes, J. J., Solway, A., Diuk, C., McGuire, J. T., Barto, A. G., Niv, Y., and Botvinick, M. M. (2011). A neural signature of hierarchical reinforcement learning. Neuron, 71(2), 370-379.

Rizzo, M. L. (2019). *Statistical computing with R*. Chapman and Hall/CRC.

Rolls, E. T., Deco, G., Huang, C. C., and Feng, J. (2022). The human orbitofrontal cortex, vmPFC, and anterior cingulate cortex effective connectome: emotion, memory, and action. Cereb Cortex.

Rung, J. M., and Madden, G. J. (2018). Experimental reductions of delay discounting and impulsive choice: A systematic review and meta-analysis. Journal of experimental psychology: General, 147(9), 1349.

Sambrook, T. D., and Goslin, J. (2015). A neural reward prediction error revealed by a meta-analysis of ERPs using great grand averages. Psychological bulletin, 141(1), 213.

Sanbonmatsu, D. M., Posavac, S. S., and Stasney, R. (1997). The subjective beliefs underlying probability overestimation. Journal of Experimental Social Psychology, 33(3), 276-295.

Schmidhuber, J. and Wiering, M. A. (1996). HQ-Learning: Discovering Markovian sub-goals for non-Markovian reinforcement learning. Technical report IDSIA-95-96, 1-13.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. Journal of neurophysiology, 80(1), 1-27.

Schultz, W. (2013). Updating dopamine reward signals. Current opinion in neurobiology, 23(2), 229-238.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. Science, 275(5306), 1593-1599.

Scott, J. (2000). Rational choice theory. Understanding contemporary society: Theories of the present, 129, 671-85.

Sears, B., Dunn, R. M., Pisklak, J. M., Spetch, M. L., and McDevitt, M. A. (2022). Good news is better than bad news, but bad news is not worse than no news. Learning and Behaviour, 1-12.

Seo, H., and Lee, D. (2007). Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game. Journal of neuroscience, 27(31), 8366-8377.

Seo, H., Kim, S., Hwang, J., and Lee, D. (2009). Valuation of uncertain and delayed rewards in primate prefrontal cortex. Neural Networks, 22(3), 294-304.

Shahnazian, D., Shulver, K., and Holroyd, C. B. (2018). Electrophysiological responses of medial prefrontal cortex to feedback at different levels of hierarchy. NeuroImage, 183, 121-131.

Sharot, T., and Sunstein, C. R. (2020). How people decide what they want to know. Nature Human Behaviour, 4(1), 14-19.

Shteingart, H., and Loewenstein, Y. (2014). Reinforcement learning and human behaviour. Current Opinion in Neurobiology, 25, 93-98.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., and Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, *362*(6419), 1140-1144.

Smith, S. L., Kindermans, P. J., Ying, C., and Le, Q. V. (2017). Don't decay the learning rate, increase the batch size. arXiv preprint arXiv:1711.00489.

Solaki, A., Berto, F., and Smets, S. (2021). The logic of fast and slow thinking. Erkenntnis, 86(3), 733-762.

Soltani, A., and Izquierdo, A. (2019). Adaptive learning under expected and unexpected uncertainty. Nature Reviews Neuroscience, 20(10), 635-644.

Solway, A., and Botvinick, M. M. (2012). Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates. Psychological review, 119(1), 120.

Spelke, E. S., and Kinzler, K. D. (2007). Core knowledge. Developmental science, 10(1), 89-96.

Spetch, M. L., Belke, T. W., Barnet, R. C., Dunn, R., and Pierce, W. D. (1990). Sub-optimal choice in a percentage-reinforcement procedure: Effects of signal condition and terminal-link length. Journal of the experimental analysis of behaviour, 53(2), 219-234.

Stagner, J. P., and Zentall, T. R. (2010). Sub-optimal choice behaviour by pigeons. Psychonomic Bulletin and Review, 17(3), 412-416.

Stephens, D. W., and Krebs, J. R. (2019). Foraging theory. In Foraging theory. Princeton university press.

Sutton, R. S., and Barto, A. G. (1998). Introduction to reinforcement learning.

Sutton, R. S., and Barto, A. G. (1999). Reinforcement learning. Journal of Cognitive Neuroscience, 11(1), 126-134.

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

Swintosky, M., Brennan, J. T., Koziel, C., Paulus, J. P., and Morrison, S. E. (2021). Sign tracking predicts sub-optimal behaviour in a rodent gambling task. Psychopharmacology, 238(9), 2645-2660.

Takahashi, Y., Schoenbaum, G., and Niv, Y. (2008). Silencing the critics: understanding the effects of cocaine sensitization on dorsolateral and ventral striatum in the context of an actor/critic model. Frontiers in neuroscience, 2, 14.

Tversky, A., and Kahneman, D. (1985). The framing of decisions and the psychology of choice. In Behavioural decision making (pp. 25-41). Springer, Boston, MA.

Valentin, V. V., Dickinson, A., and O'Doherty, J. P. (2007). Determining the neural substrates of goal-directed learning in the human brain. Journal of Neuroscience, 27(15), 4019-4026.

Van de Vijver, I., Ridderinkhof, K. R., and Cohen, M. X. (2011). Frontal oscillatory dynamics predict feedback learning and action adjustment. Journal of cognitive neuroscience, 23(12), 4106-4121.

Von Neumann, J., and Morgenstern, O. (2007). Theory of games and economic behaviour. In Theory of games and economic behaviour. Princeton university press.

Vonder Haar, C., Ferland, J. M. N., Kaur, S., Riparip, L. K., Rosi, S., and Winstanley, C. A. (2019). Cocaine self-administration is increased after frontal traumatic brain injury and associated with neuroinflammation. European journal of neuroscience, 50(3), 2134-2145.

Walton, M. E., Bannerman, D. M., and Rushworth, M. F. (2002). The role of rat medial frontal cortex in effort-based decision making. Journal of Neuroscience, 22(24), 10996-11003.

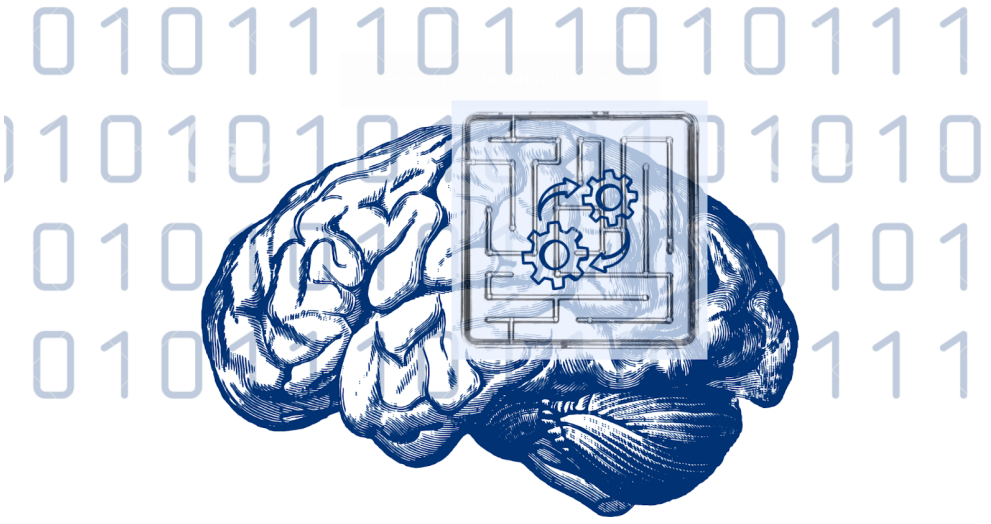
Weiskopf, N., Hutton, C., Josephs, O., and Deichmann, R. (2006). Optimal EPI parameters for reduction of susceptibility-induced BOLD sensitivity losses: a whole-brain analysis at 3 T and 1.5 T. Neuroimage, 33(2), 493-504.Wickens, C. D., and McCarley, J. S. (2019). *Applied attention theory*. CRC press.

Zentall, T. R. and Case, J. P. (2018). Sub-optimal choice in pigeons: Does the predictive value of the conditioned reinforcer alone determine choice?. Behavioural Processes, 157, 320-326.

Zentall, T. R. (2016). Resolving the paradox of sub-optimal choice. Journal of Experimental Psychology: Animal Learning and Cognition, 42(1), 1.

Zentall, T. R., and Stagner, J. (2011). Maladaptive choice behaviour by pigeons: an animal analogue and possible mechanism for gambling (sub-optimal human decision-making behaviour). Proceedings of the Royal Society B: Biological Sciences, 278(1709), 1203-1208.

Zhang, D., and Gu, R. (2018). Behavioural preference in sequential decision-making and its association with anxiety. Human brain mapping, 39(6), 2482-2499.



10. Supplementary Material

10. Supplementary Material.

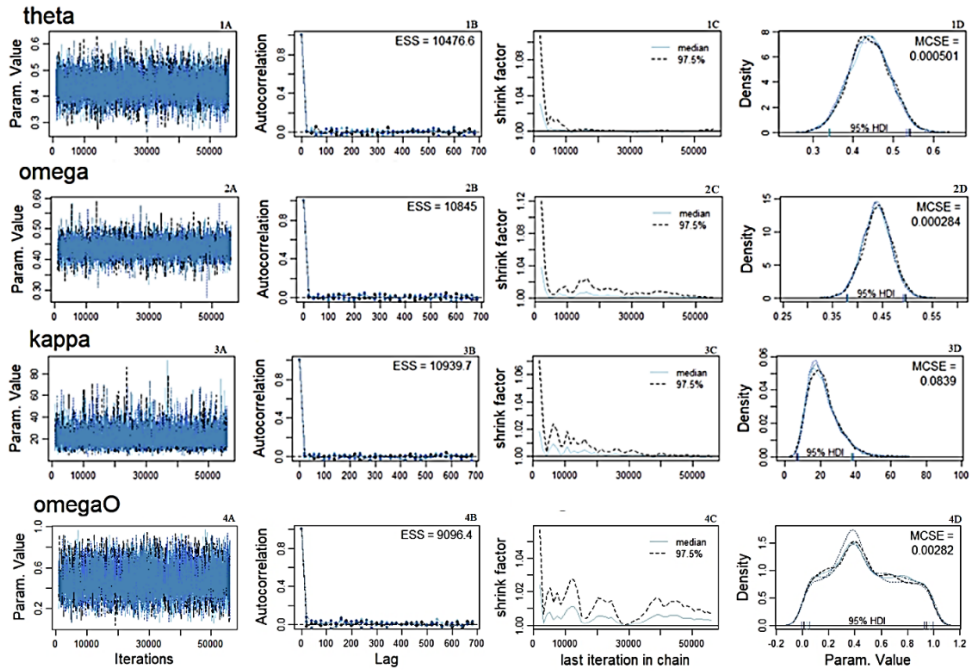
Feasibility of Hierarchical Model in Experiment 1

The values for parameters theta (SD = .05; MCSE = .0005; EES = 10652.8), omega g (SD = .029; MCSE = .0002; EES = 11222.9) and omega 0 (SD= 0.08; ESS = 9096.4; MCSE= .002) were credible and not biased by the initial values of the Markovian chains. Besides, the variance between the chains concerning the variance within the chains showed that MCMC converged well enough. The visual and numerical assessment of MCMC showed that the parameters were representative and stable and converged throughout the range of the retrieval (burn-in period) until reaching the posterior distribution. In addition, Supplementary Figure S1 shows that the shrink factor of the model was below 1.1, indicating a clear convergence of the model (Gelman and Rubin, 1992; Rizzo, 2019) and high autocorrelation function (ACF), showing that the estimated values for the parameters are stable among the chains' steps. In order to compute the whole effect of the ACF across the model, the number of steps divided by the sum of each ACF compiles an effective sample size indicator (ESS). Values around 10000 indicate the efficiency of the model and correspond with a small Markovian Chain Standard Error (MCSE). In the case of the parameters theta (ESS = 10476.6; MCSE= .0005), omega (ESS = 10845; MCSE = .0003) and kappa (ESS = 10939.7 ; MCSE= .008) the numerical checks supported the reliability of the model.

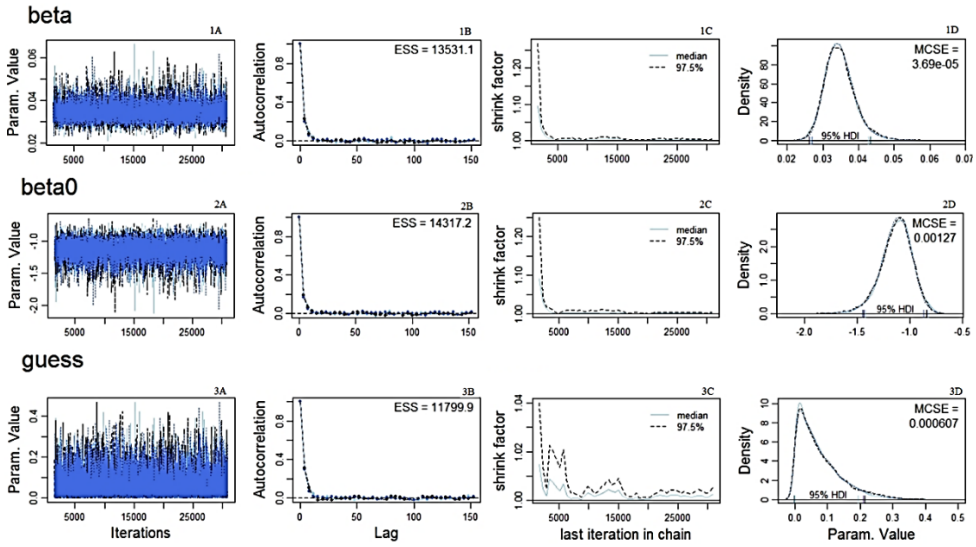
Feasibility of Hierarchical Model in Experiment 2

The values for parameters B0 (MCSE = .00226; EES =11485), B1 (MCSE = .0006; EES =8843) and "guessing coefficient" (MCSE = .0007; EES =8506.1) are credible. In addition, the shrink factor (Gelman and Rubin, 1992) for every predictor was below 1.1. These parameters are representative, stable and converge till the posterior distribution (**Supplementary Figure S2**). The shrink factor was below 1.1 indicating good accuracy of the model. In addition the parameters Beta1

(ESS = 13531.1; MCSE= .0004), Beta0 (ESS = 14317.2; MCSE = .0001) and guess (ESS = 11799.9; MCSE=.0006) also supported the reliability of the model.



Supplementary Figure 11.1. Representativeness and accuracy of model 1. The evaluation of the model for experiment 1 can be focused on two dimensions: visual examination and numerical description. The left panels (1A, 2A, and 3A) are trace plots that follow the superimposition of MCMC along with the iterations or burn-in period. Regarding the graphs 1A, 2A, and 3A, the parameters theta, beta, and Kappa converged after the Markovian Chains were launched. When the chains overlap across the iterations means, they are fully exploring the range of the data distribution and that the initial values are not misleading such trajectories. Considering the density plots in the panels 1D, 2D and 3D, for most of the cases, the data coincides with the 95 % credible intervals HDI of the posterior densities. These density plots compare three random samples of chains and display them into smoothed histograms delimitated by an individual HDI. For the three parameters considered in the model, the three sample chains mostly overlay each other.



Supplementary Figure 11.2 Representativeness and accuracy of model 2. The evaluation of the model for experiment 2, regarding a visual examination and numerical description. The left panels (1A, 2A, and 3A) are trace plots for the parameters Beta1, Beta0, and Guessing factor. These plots show convergence for MCMC after the Markovian Chains are retrieved. For most of the cases, considering the density plots in the graphs 1D, 2D, and 3D, the data is within the 95 % credible intervals HDI of the posterior densities. For all the parameters considered in the model, the three sample chains mostly overlay each other.

11. Table of Figures

UNIVERSITAT DE BARCELONA

Cognition and Brain Plasticity Unit