Electrophysiological correlates of reinforcement learning and credit assignment

Inaugural-Dissertation zur Erlangung des Doktorgrades
der Philosophisch-Pädagogischen Fakultät

der

Katholischen Universität Eichstätt-Ingolstadt

vorgelegt von

Franz Wurm

2020

Referent:	Prof. Dr. Marco Steinhauser
Korreferent:	Prof. Dr. Michael Zehetleitner
Tag der Disputation:	29.09.2020

Abstract

Reinforcement learning constitutes a valuable framework for reward-based decision making in humans, as it breaks down learning into a few computational steps. These computations are embedded in a task representation that links together stimuli, actions, and outcomes, and an internal model that derives contingencies from explicit knowledge. Although research on reinforcement learning has already greatly advanced our insights into the brain, there remain many open questions regarding the interaction between reinforcement learning, task representations, and internal models. Through the combination of computational modelling, experimental manipulation, and electrophysiological recording, the three studies of this thesis aim to elucidate how task representations and internal models are shaped and how they affect reinforcement learning. In Study 1, the manipulation of action-outcome contingencies in a simple one-stage decision task allowed to investigate the impact of explicit knowledge about task learnability on reinforcement learning. The results highlight the flexible adjustment of internal models and the suppression of central computations of reinforcement learning when a task is represented as not learnable. Using a similar manipulation, Study 2 investigates how this influence of explicit knowledge on reinforcement learning holds under the increasing complexity of a two-stage environment. Again, pronounced neural differences between task conditions indicate separable computations of reinforcement learning and, more importantly, the selective influence of explicit knowledge and internal models on reinforcement learning. Study 3 uses a novel task design which necessitates inference about plausible action-outcome mappings, and thus, credit assignment. The findings suggest that multiple task representations are neurally conceptualized and compete for action selection, thereby solving the structural credit assignment problem. In sum, the studies of this thesis highlight the importance of reinforcement learning as a central biological principle and draw attention to the necessity of flexible interactions between reinforcement learning, task representations, and internal models to cope with the varying demands from the environment.

Contents

Introduction

If scientific inquiry has taught us one thing, it is that the world and its mechanisms are highly complex and almost always more complicated than we initially considered. Examples for this escalating complexity are ubiquitous across multiple domains, from theoretical physics to economics. Thinking about complex systems quickly leads to the generation of models which reduce the complexity and abstract away unnecessary detail. While this statement is true for almost every field of science, in cognitive neuroscience, where the object of investigation is the human brain, the word "model" applies in a twofold meaning. Since the cognitive revolution in the 1950's (Miller, 2003), one of the most prevailing and arguably most fruitful theories describes the mind as an information processing system (A. Newell & Simon, 1972; Norman, 1976; Simon, 1978). The simplifying idea behind this concept is that the brain, similar, for example, to a computer or a calculator, takes a certain input and formulates a certain output based on some internal computation. The field of computational cognitive neuroscience is specifically puzzled with the question on the qualitative and quantitative computations of the brain and seeks to find a formalization thereof (Kriegeskorte & Douglas, 2018). One of the currently most prominent theories, that allows such a input-output formalization for the brain, is reinforcement learning, as it quantitively conceptualizes the brain as an agent which performs actions based on past experience to maximize future reward (Dayan & Niv, 2008; Niv, 2009; Niv & Langdon, 2016; Sutton & Barto, 2018). This is the first meaning of the word "model" which I call a computational model and which cognitive neuroscience shares with other fields of research. The central aspect of a computational model is that the scientist formulates hypotheses about its object of inquiry. However, based on the Helmholtzian perceptive (Dayan, Hinton, Neal, & Zemel, 1995; von Helmholtz, 1909), the brain as the object of neuroscientific inquiry is regarded as a model or hypothesis generator itself and its main function is to make inference about the probable causes of its inputs. This resulting model or hypothesis is the second meaning of the word

"model" which I call an internal model, and which is unique to cognitive neuroscience (and arguably psychology). The central aspect of an internal model is that the human brain as the object of investigation formulates hypotheses about its own object of investigation.

The following empirical studies operate at the cross-section between both computational and internal models. My goal is to advance our insights on how the internal model (and the representation of the environment) can be accounted for by computational models of reinforcement learning. In this endeavor I carried out a series of three studies in which human participants performed sequential decisions to obtain reward. As a central manipulation in all three experiments the task complexity emphasized the requirement of a flexible representation of the task structure. First, I ask, how explicit knowledge about the causal structure of the environment manifests on the neural level and how different internal models impact on reinforcement learning during feedback processing. Second, I ask how this influence of the internal model on reinforcement learning changes as a function of task complexity. Third, I ask, how the experience can be used for inference and the arbitration between different plausible representations of the causal structure in the environment. I attempt to contribute to answering these questions through the application of experimental manipulation, analysis of electrophysiological data (EEG) and the implementation of computational models. In combination, these methods allow to draw a detailed picture of the human brain as a complex but efficient reinforcement learning agent which uses representations of the environment to infer inputs and maximize outputs.

A computational model of reinforcement learning

Models of reinforcement learning constitutes a well-defined set of algorithms that formalizes in a normative way how an agent learns to choose between different actions in order to maximize rewards and minimize punishment (Sutton & Barto, 2018). In the scope of

my study, reinforcement learning poses the central framework for the investigation of decision making in the human brain.

The fundamental appeal of reinforcement learning arises from its ability to solve decision and learning problems using a simple and straightforward computational methodology. By breaking down complex problems into a few central concepts and ideas, computational models of reinforcement learning achieve great precision in finding close to optimal solutions. At its core, the idea behind reinforcement learning is that learning about what to do derives from interaction with the environment (Sutton & Barto, 2018). In a formalized way, this interaction follows the notation: The agent finds itself in a specific *state* (*s*) and takes a particular *action* (*a*). He subsequently observes an *outcome* (*r*) based on the previous action and finds himself in a new state which again might require another action leading to a new outcome and so forth. Consider, for example, a thirsty agent. In this state it can take different actions, such as eating, drinking or continue reading. The agent decides to eat something, but as you can imagine, the outcome is not as rewarding as expected, leaving the agent in a state of thirst (although maybe not hungry anymore) until it decides to take the action of drinking something. What the agent may have learned in this situation, is that only drinking in contrast to eating leads to the desired outcome that is moving away from a state of thirst. Therefore, if it ever experiences thirst again it may take the drinking action first.

In an experimental setup, similar reward-based decision problems are usually studied in the so-called bandit task, drawing on its analogy to a slot machine or one-armed bandit. Bandit problems have been extensively studied in machine learning (Berry & Fristedt, 1986; Kaelbling, Littman, & Moore, 1996; Macready & Wolpert, 1998; Sutton & Barto, 2018) and cognitive science (J. D. Cohen, McClure, & Yu, 2007; Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Frank, Seeberger, & O'Reilly, 2004), as they offer a simplified and controllable environment for the investigation of decision making. On each trial of the task an agent is faced with a decision between multiple alternative actions, each of which is associated with an

outcome drawn from a fixed but initially unknown probability distribution. Usually the outcome consists of some form of monetary payoff motivating agents to choose actions which maximize the total payoff over a sequence of trials. While multiple computational models for optimal reward-based decision-making have been proposed for solving the bandit task (Lee, Zhang, Munro, & Steyvers, 2011; Steyvers, Lee, & Wagenmakers, 2009), I focus on the reinforcement learning model, as it is most relevant in the scope of this thesis.

First, I will take a closer look at the underlying mathematical computations and latent variables of reinforcement learning that allow an agent to maximize payoff by connecting the ideas about states, actions and outcomes encountered previously. I illustrate how the central latent variables of reinforcement are calculated using the bandit problem as introduced above. Based on the contingencies of the task, each action in the bandit task is followed by an outcome with a specific value. It is this value that action selection should have been based on. If an agent were informed on the values associated with his actions prior to taking the actions, solving the bandit task would be trivial: always (greedily) selecting the action with the highest value maximizes the total payoff. Although some studies provided participants with the exact action values (e.g. Li, Delgado, & Phelps, 2011; Walsh & Anderson, 2011), action selection can usually only be based on an *estimated value of an action a* in the given state *s*, denoted as *Q(s,a)*. The goal of reinforcement learning is to optimize action selection by bringing the estimated value of an action as close to the exact (i.e., to be obtained) value of that action as possible. It does so by utilizing two simple yet powerful calculation steps. First, after observing the action-dependent outcome, reinforcement learning calculates the *prediction error* δ[1] following the equation

[1] Technically, the term "temporal difference error", originally proposed by Sutton and Barto (2018) as an extension of the "prediction error" proposed in the Rescorla-Wagner model (Rescorla & Wagner, 1972) would be correct. Besides the difference between the observed and expected value of an action, the temporal difference error also incorporates the summed expected rewards for all states observed in the future. However, this term is mostly dropped in the neuroscientific literature. See Sutton & Barto (2018) for further reading and the formal notation.

$$\delta(\mathrm{t}) = [r(\mathrm{t}) - Q(\mathrm{a}, \mathrm{s}, \mathrm{t})], \quad (1)$$

where *r(t)* denotes the observed outcome received in that trial and *Q(a,s,t)* the expected value of the chosen action in a specific state. In the second step, this prediction error is then used to update the action values according to the equation

$$Q(a, s, t+1) = Q(a, s, t) + \alpha * \delta(t), \quad (2)$$

where α is the learning rate, which controls for the speed of updating of new information incorporated in the prediction error. Calculating these steps incrementally, that is again for every trial of the bandit task, leads to an approximation of the exact values of the different actions. Action selection on the next trial is then determined by transferring the updated action values into action probabilities using the softmax function

$$P(a_t = a \mid s) = \frac{exp\ (\beta * Q(a, s, t))}{\sum_{a'} exp\ (\beta * Q(a', s, t))}, \quad (3)$$

where the inverse temperature β guides the stochasticity of the choices. Put simply, the softmax function is a normalizing procedure, which determines how strongly the differences in action values are translated into action probabilities. Taken together, the calculation of a prediction error, updating of the expected action value, and action selection are the computational basis for reinforcement learning and in the long run this three-step procedure guarantees close to optimal decision making which maximizes mean payoffs for an agent.

Historically, the computational model of reinforcement learning is preceded by a wealth of animal studies investigating trial and error learning (Sutton & Barto, 2018). Already in 1898, Edward Thorndike noted the impact of reward on subsequent behavior. In his experiments, cats were placed in puzzle boxes and food was presented outside of the boxes. The boxes contained a setup of different levers and strings, which, when activated in the correct order, released the cat, and allowed access to the food. Thorndike observed that his subjects not only were able to escape from the boxes but also became progressively faster.

Derived from this observation he introduced the “Law of Effect”. The law of effect states that “the greater the satisfaction or discomfort, the greater the strengthening or weakening of the bond” (Thorndike, 1911, p.244). “Bond” in his words referred to the link or association between a state of the environment and an action of the agents. In other words, action values and action probabilities increase if an action is followed by a reward but decrease after a loss. Although the law of effect is devoid of a mathematical formalization, it is almost synonymous with reinforcement learning. However, Thorndike was not the only to notice the connection between reward or punishment and subsequent decision making. In the 1927 translation of the work done by the now famous Russian physiologist Ivan Pavlov in the late 19th century, the term “reinforcement” was introduced to describe the strengthening of a behavioral pattern which resulted from the delivery of a rewarding stimulus, i.e., a reinforcer. Crucially, Pavlov noted that given an appropriate temporal relationship between a rewarding stimulus and another stimulus or even an action so-called “conditioned reflexes” emerge. While studying the digestive system in dogs, he observed that innate responses (reflexes) to specific triggering stimuli can also be elicited by other, quite unrelated stimuli if they were paired with the initial triggering stimulus. More explicitly, in his experiment dogs did start salivating shortly after being presented with food. However, after several repetitions of pairing the saliva-eliciting food with the sound of a metronome, his test subjects already began salivating when only the sound of the metronome was presented. Establishing such novel connections is now called Pavlovian or classical conditioning. In contrast to Pavlov’s classical conditioning, which emphasized the association between conditional and unconditional stimuli (sound and food), work in line with Thorndike’s law of effect emphasized that learning depends on the consequences of subsequent behavior and was called instrumental or operant conditioning. Please note, however, that both forms can be subsumed under der framework of temporal-difference learning, as described in more detail elsewhere (Sutton & Barto, 2018). Central to reinforcement learning is the agent’s ability to learn a prediction of the environment and to

adapt behavior in anticipation of an expected outcome. This could either be the prediction of a stimulus or state signaling subsequent outcomes (classical conditioning) or that of an action leading to an outcome (instrumental conditioning).

The neural correlates of reinforcement learning

Early research on conditioning in animals was exclusively based on behavioral measures (e.g. response rates and reaction times). However, a new era of reinforcement learning started in the late 20th century as evidence accumulated that the latent variables of reinforcement learning (i.e. prediction errors and action values) manifest on the neural level (for a recent review see Schultz, 2016). While studying the relationship between the neurotransmitter dopamine and muscle activity in monkeys, it was discovered that phasic bursts of dopamine initially associated with the delivery of a specific stimulus (in this case a rewarding stimulus, i.e. food) shifted in time over the course of the training procedure (Romo & Schultz, 1990; Schultz & Romo, 1992). As with the shift of the salivation onset for Pavlov's dogs from food delivery to the predictive sound of the metronome, the observed pattern of neural firing revealed classical conditioning. This finding paved the way for the still influential hypothesis, that reward prediction errors are reflected in the neural activity of dopamine neurons (Montague, Dayan, & Sejnowski, 1996; for reviews see Glimcher, 2011; Schultz, 2016). This theory proposes that phasic dopamine activity from neurons in the ventral tegmental area (VTA) and the substantia nigra (SNpc), which are the main sources of dopamine in the mammalian brain, act as a teaching signal for target areas, which receive projections from these midbrain structures. Crucially, dopamine was hypothesized to reflect a reward prediction error, i.e., the difference between an expected and an observed outcome, thereby carrying information as provided by a reinforcement learning agent. While it was known at that time, that dopamine is associated with reward (Olds & Milner, 1954), the link

to formalized reinforcement learning, as developed in the growing field of machine learning, was astonishing. In line with their prediction, Schultz and colleagues found strong resemblance when comparing the prediction errors of a reinforcement learning model of classical conditioning with the phasic activity of dopamine neurons: A negative prediction error was mirrored by a decrease in phasic dopamine activity and a positive prediction error was mirrored by an increase in phasic dopamine activity (Montague et al., 1996; Schultz, Dayan, & Montague, 1997). Since these pioneer studies, the characteristic signatures of prediction errors in phasic dopaminergic responses have been replicated in many different settings (e.g. Bayer & Glimcher, 2005; Hollerman & Schultz, 1998; Takikawa, Kawagoe, & Hikosaka, 2004; Tobler, Dickinson, & Schultz, 2003) and the quantitative implications of the prediction error hypothesis, which lie deep within the theoretical foundations of reinforcement learning, have been reliably found to be reflected in dopamine responses (e.g. Bayer, Lau, & Glimcher, 2007; Fiorillo, Tobler, & Schultz, 2003; Roesch, Calu, & Schoenbaum, 2007; for a review, see Niv, 2009).

However, these studies almost exclusively relied on intracranial methods in animals. While these methods allow for both a good temporal and spatial resolution, they are highly invasive. Although conditioning in humans and animals can be regarded as similar (Niv, 2009), the goal for many researchers is to give a comprehensive account of human decision making. The most prominent technique for studying reinforcement learning in the human brain is functional magnetic resonance imaging (fMRI), as it allows to scan the brain in a non-invasive way. Key advantage of this method is that it is particularly suited to identify spatial patterns of brain activations. Besides a low signal-to-noise ratio and a poor temporal resolution, the major disadvantage of fMRI is its challenge for statistical analyses of the extensive data (Gazzaniga, Ivry, & Mangun, 2014). Crucially, computational models are able to counteract this problem as they allow to search for latent variables which gave rise to the observed data. More specifically, the precise quantification of prediction errors and action

values within models of reinforcement learning enable to track the associated processes in the brain. This is achieved by searching for the dynamics of the latent variables in the brain data and identifying significant clusters of overlapping patterns between the model and the data (for a review of this method see Gläscher & O'Doherty, 2010). While early fMRI studies did not exploit the advantages of the analyses described above (Berns, McClure, Pagnoni, & Montague, 2001; Knutson, Adams, Fong, & Hommer, 2001; Pagnoni, Zink, Montague, & Berns, 2002), latent-variables analyses have become an indispensable tool for research on the connection between reinforcement learning and human brain activity. For example, in a seminal study latent-variable analyses revealed that prediction errors signals are dissociated between the dorsal and ventral striatum dependent on whether the task involved active decision making (instrumental conditioning) or not (Pavlovian conditioning) (O'Doherty et al., 2004). Finding these significant correlations not only supports the computational model used for the generation of the predictions but also greatly advances our knowledge about the architectural structure of the brain. However, one has to keep in mind, that the origin of blood oxygen level dependent (BOLD) response, which is quantified in fMRI studies, is still far from being fully understood (Turner, 2016). Therefore, the application of multiple different measures is advisable (O'Doherty, Hampton, & Kim, 2007).

Another technique which is extensively used for studying reinforcement learning in the human brain is EEG. In contrast to fMRI, EEG provides a good means to identify and separate the temporal patterns underlying the diverse cognitive processes in the brain. Unsurprisingly, there exists a wealth of EEG studies that investigates the brain activity associated with reinforcement learning. The two components of the human event-related potential (ERP) most strongly associated with processing of outcome and reward are the feedback-related negativity (FRN; for reviews see San Martín, 2012; Walsh & Anderson, 2012; Sambrook & Goslin, 2015; Proudfit, 2015; Holroyd & Umemoto, 2016; Krigolson,

2017; Cockburn & Holroyd, 2018) and the P3 (for reviews see Polich, 2007, 2020; San Martín, 2012).

The FRN occurs at frontocentral electrode sites around 250 ms after the presentation of feedback and manifests as a strong negative deflection following losses compared to wins. It was first reported in a task in which participants received visual feedback about the correctness of a time estimation (Miltner, Braun, & Coles, 1997), but has since been replicated in multiple experimental settings. It is reportedly modulated by multiple outcome dimensions, such as valence, reward magnitude, and probability (for a review and discussion of inconsistent findings, see San Martín, 2012; Walsh & Anderson, 2012). Most importantly, this component has been claimed to reflect a reward prediction error as calculated by a temporal difference reinforcement learning algorithm (Holroyd & Coles, 2002). Although this claim has been investigated in many studies, most of these studies lack the powerful latent-variable analyses approach described for the fMRI research in reinforcement learning. However, a meta-analysis revealed that this claim is consistent with the data and the FRN reflects a reward prediction error (Sambrook & Goslin, 2015). Recent evidence from studies which harness the advantages of computational modelling via latent-variable analyses further supports this idea (e.g. Chase, Swainson, Durham, Benham, & Cools, 2011; Walsh & Anderson, 2011; Sambrook, Hardwick, Wills, & Goslin, 2018).

In short temporal succession to the FRN, the P3 occurs at posterior electrode sites around 300-600 ms after the presentation of the outcome. The P3 has been extensively investigated using the so-called “oddball” paradigm (e.g. Duncan-Johnson & Donchin, 1977; Pritchard, 1981), in which a low-frequency target stimulus (i.e., the oddball) is merged into a stream of high-frequency but non-target stimuli. Similar to the FRN, the P3 is also implicated to be modulated by distinct outcome dimensions, such as valence, magnitude, probability and need for behavioral adaptation (for a review and discussion, see San Martín, 2012). However, in contrast to the FRN which is suggested to foremost reflect a first evaluation of primary

outcome attributes (e.g. win vs. loss), the P3 is suggested to reflect a more in-depth evaluation of secondary attributes (e.g. magnitude or probability) (Bernat, Nelson, & Baskin-Sommers, 2015; Cavanagh, 2015; Hajcak, Moser, Holroyd, & Simons, 2006; Nieuwenhuis, Holroyd, Mol, & Coles, 2004). Besides the obvious connection to measures of surprise (Kolossa, Fingscheidt, Wessel, & Kopp, 2012; Kolossa, Kopp, & Fingscheidt, 2015; Kopp et al., 2016; Mars et al., 2008; Seer, Lange, Boos, Dengler, & Kopp, 2016), the P3 has also been implicated in reflecting learning (Fischer & Ullsperger, 2013). Crucially, this link even holds after controlling for surprise (Jepma et al., 2018, 2016; Nassar, Bruckner, & Frank, 2019).

Taken together, the FRN and P3 have been suggested to reflect distinct processes of feedback processing. For the scope of this thesis, this distinction is of special interest regarding reinforcement learning where the FRN is suggested to reflect a reward prediction error (Holroyd & Coles, 2002; Sambrook & Goslin, 2015), whereas the P3 seems to reflect the value of actions (Fischer & Ullsperger, 2013).

The credit assignment problem

The idea that the brain acts in accordance with reinforcement learning principles is supported by the findings from different lines of research and the hypothesis that phasic dopamine reflects reward prediction errors is broadly accepted. Yet, this picture is overly optimistic. For example, the observable structure of the extensively studied two-armed bandit task, where one action is followed by one outcome, implies a parsimonious and simple representation of the central contingency between the input (i.e. states and outcomes) as well as the possible outputs (i.e. actions) for the agent. Although such a reductionist approach allows for a precise investigation of decision making in the laboratory, the bandit task (as do most experimental setups) lacks ecological validity. Hence the generalization of findings from

simple computational models to account for decision making in naturalistic settings can be biased (Nassar & Frank, 2016). Equivalent limitations of the action and outcome space as in the bandit task are only rarely found in naturalistic settings, where a vast amount of actions can be made, and where rewards are often delayed in time. While the biological agents (e.g., animals and humans) perform well in naturalistic environments, artificial agents (e.g., a computational model of reinforcement learning) quickly break down and fail to work efficiently when faced with complex environments. This conundrum has been termed the "curse of dimensionality" (Bellman, 1957) because the number of parameters that must be represented and learned grows exponentially with the size of the task structure. This challenge resonates well with what has been introduced as the so-called "credit assignment problem" (Minsky, 1961). The credit assignment problem asks how agents assign credit for a reward to the several actions which may have been involved in producing it, or, put differently, how agents form a representation of the causal structure between actions and outcomes. More specifically, the credit assignment problem can be subdivided into a temporal and a structural aspect.

Already the early researchers on classical and instrumental conditioning were aware of the temporal aspect of the credit assignment problem. Stimuli or actions in the past must receive credit or blame for the following consequences and therefore the formation of a predictive association required some sort of backward effect, sometimes called "spread of effect" (Thorndike, 1933). On the mathematical level, this backward effect of reinforcement learning is implemented in the updating of the expected action value using the prediction error (Equation 2). Already Pavlov (1927) pointed out that the existence of transient traces in the nervous system should allow learning (or conditioning) and this idea was incorporated in Hull's influential learning theory to account for a variety of findings on instrumental conditioning (Hull, 1932, 1942). It was proposed that stimulus traces can bridge the gap in time between actions and consequent reward, making the action eligible for modification thus

allowing the updating of its action value. Although highly speculative at this time, the idea of eligibility traces for solving the temporal credit assignment problem was quickly incorporated as a basic mechanism in the computational model of reinforcement learning (Sutton & Barto, 2018) and has received ample evidence as a biological plausible mechanism (Roelfsema & Holtmaat, 2018). Learning with eligibility traces is highly efficient even in complex environments because the backward effect of prediction errors is enhanced which subsequently increases the update of action values thus speeding up learning.

In addition, the credit assignment problem also has a strong structural component. Besides the need for bridging the time gap between an action and an outcome, agents are often required to identify the relevant elements of the task but ignore irrelevant elements (i.e., stimuli, actions, outcomes). An important role for the solution of this structural credit assignment problem is attributed to the prefrontal cortex (Asaad, Lauro, Perge, & Eskandar, 2017; Jocham et al., 2016; Noonan, Chau, Rushworth, & Fellows, 2017; Noonan et al., 2010). For example, Noonan and colleagues (2010) showed that the orbitofrontal cortex (OFC), a region strongly implicated in reward-based learning and decision making (FitzGerald, Seymour, & Dolan, 2009; Price, 2007), was causally involved in structural credit assignment. In their study, monkeys were trained to perform a bandit task, that on every trial required a decision between different actions followed by an outcome. In comparison to healthy monkeys, animals with OFC lesions were no longer able to learn the contingencies between actions and outcomes. The law of effect was apparently no longer effective as the lesioned monkeys lacked the ability to perform actions that maximize outcome. Interestingly, the backward-effect of reinforcement learning still was effective but apparently spilled across the trial structure, so that previous actions irrelevant for the present outcome were assigned credit and hence were updated. This suggests that while the structural credit assignment could not be solved, temporal credit assignment was still intact. In a recent study by the same group, these findings were extended to human subjects (Noonan et al., 2017). Again, only humans with

OFC lesions had difficulties in learning the representation of the simple bandit task. Although it has been proposed that dynamic interactions between the subregions of the PFC support structural credit assignment (Stolyarova, 2018) the underlying computational model and algorithm is still a matter of debate.

In conclusion, reinforcement learning is a powerful computational framework for understanding decision making. However, in complex environments, the credit assignment problem arises because of the temporal delays and structural ambiguities between actions and outcomes. Solving this temporal and structural credit assignment problem is essential for reinforcement learning agents in order to optimize decision making. Although computational and biologically plausible solutions have been postulated for the temporal credit assignment problem, the exact implementation of its structural counterpart within the brain is still being discussed.

Multiple systems of reinforcement learning

So far, I have argued that the brain can be regarded as an agent which bases decision making on the reinforcement learning principle to maximize reward and minimize punishment in a given task. However, in contrast to animals and humans, the computational model of reinforcement learning quickly runs into credit assignment problems when faced with ambiguous or complex task structures. Moreover, it has been assumed that humans and animals possess multiple parallel reinforcement learning systems that compete for action selection (Daw, Niv, & Dayan, 2005; Dickinson & Balleine, 2002).

For example, a common taxonomy of instrumental behavior, i.e., behavior that is shaped by the law of effect, distinguishes between goal-directed and habitual learning (Balleine & Dickinson, 1998; Dickinson & Balleine, 1995, 2002; Graybiel, 2008). Following this distinction, habitual behavior is acquired by repeatedly performing an action in a given

context (i.e. stimulus-response learning), whereas goal-directed behavior is acquired by performing an action to obtain an outcome (i.e. response-outcome learning). Formally, instrumental behavior is considered to be goal-directed if it meets two criteria (Dolan & Dayan, 2013). First, behavior must reflect knowledge about the link between cause and effect. Second, the outcome should be motivationally significant, that is attractive or worthwhile during action selection. In contrast, habitual behavior is hypothesized to be merely "stamped in" by the history of past reinforcement, disconnecting it from the current value of an action. Experimentally, the dichotomy between goal-directed and habitual behavior is usually demonstrated using a devaluation procedure. After learning of an action-outcome association (e.g. which arm is better in a bandit task), the outcome is devaluated, that is its delivery has no more rewarding property (i.e. satiety after food). Under goal-directed control, an agent should quickly cease to act in accordance with the previously learned action-outcome association, showing that behavior is governed by a representation of the outcome (Adams & Dickinson, 1981; Gillan, Otto, Phelps, & Daw, 2015). Under habitual control, an agent should continue to act as usual, even when the outcome is undesirable. Interestingly, behavior can be regarded as a mixture of both modes of control with multiple contextual and intraindividual variables such as duration of training, task complexity, working memory capacity and working memory load (Adams, 1982; Kool, Gershman, & Cushman, 2017, 2018; Otto, Gershman, Markman, & Daw, 2013; Otto, Raio, Chiang, Phelps, & Daw, 2013) affecting the trade-off between them. Furthermore, lesion studies in rats implicate that both systems are dependent on distinct neural networks in the brain, including various parts of the frontal cortex and striatum (for a review see Daw & O'Doherty, 2014). However, findings in healthy animals support the idea of a dynamic interaction and dependency between both modes of control (Wassum, Cely, Maidment, & Balleine, 2009).

Already in 1948, Edward Tolman argued, that the law of effect (i.e. the acquisition of instrumental behavior) is insufficient to account for all forms of mammalian learning.

Evidence for this statement was drawn from latent learning experiments which asked if learning is possible even in the absence of reward. In the earliest version of this experiment (Blodgett, 1929), two groups of rats were placed in a labyrinth. On the one hand, the reward group ran the maze and always found a desirable outcome at a goal position. On the other hand, the no-reward group initially ran the maze without such an outcome. As expected, the rats in the reward group showed learning, as indicated by their increasing performance over time. Crucially, while the rats in the no-reward group did not show any signs of instrumental behavior during the initial unrewarded phase, they quickly caught up with the performance in the reward group, when an outcome was suddenly introduced at the goal position. This demonstrated that the rats in the no-reward group acquired knowledge about the structure of the maze, which later facilitated learning when the reward was introduced. In conclusion, it was assumed that the experimental setup unmasked the existence of latent learning (Blodgett, 1929). In his now classic work, Tolman argued that this and similar findings propagate the existence of so-called cognitive maps, i.e., mental representations of the environment in which the agent operates (Tolman, 1948). The idea of cognitive maps has strongly influenced the field of cognitive neuroscience (Dolan & Dayan, 2013). A central brain structure which supports the notion of cognitive maps quite literally is the hippocampus. Most famously, the hippocampus was found to consist of so-called place cells, which provide a neural representation of the rat's environment (O'Keefe & Nadel, 1978) and seem to be activated according to an internal exploration or planning in the future (Johnson & Redish, 2007; Pfeiffer & Foster, 2013; van der Meer & Redish, 2009). Besides hippocampal areas, multiple other areas such as the prefrontal cortex, the amygdala and the dorsomedial striatum have been implicated in the representation of cognitive maps (Balleine, 2005; Balleine & Dickinson, 1998; Corbit & Balleine, 2003; Yin, Ostlund, Knowlton, & Balleine, 2005).

Crucially, the notion of cognitive maps also had strong implications on the study of reinforcement learning in the brain, leading to the distinction between model-free and model-

based reinforcement learning. On the one hand, model-free reinforcement learning is solely driven by reward and prediction errors to estimate the contingencies in the world. Most of the literature and research reported so far deal with this aspect of reinforcement learning. On the other hand, model-based reinforcement learning additionally bases decision making on an internal model of the world which (pre)assigns contingencies between states, actions, and outcomes. Model-free and model-based behavior can be functionally characterized in a mutually exclusive way. On the one hand, model-free behavior is supposed to be automatic, computationally efficient, and inflexible, whereas model-based behavior can include active deliberation, is computationally costly but allows flexible adaptation to changing task contingencies.

The emerging dichotomy between model-free and model-based control and its mathematical formalization has sparked a variety of novel paradigms which allow to contrast both modes of reinforcement learning. Most famous is the widely used sequential two-choice Markov decision task (Daw, Gershman, Seymour, Dayan, & Dolan, 2011). As an extension of the standard bandit task, agents are faced with a choice between two stimuli at a first-stage decision state. Based on a probabilistic transition structure, each action at this first-stage decision state is followed by one of two second-stage decision states. More specifically, each first-stage action is commonly (70%) followed by one associated second-stage state and only rarely (30%) with the other second-stage state. At the second-stage state, the participants are again faced with a choice between two stimuli with each second-stage state being associated with a different stimulus pair. Finally, each action is followed by a binary outcome. The probability for a positive outcome for each of the four actions at the second-stage decision states is slowly (and independently) changing throughout a block, following a Gaussian random walk. Critically, flexible decision making in this task is dependent on an internal model which accounts for both the probabilistic but fixed contingencies between first-stage and second-stage states and the probabilistic but volatile contingencies between second-stage

actions and outcomes. While model-free learning is not equipped with such an internal model and hence action selection (at the first-stage state) is only sensitive to previous outcomes, model-based learning exploits the explicit knowledge about the transition structure from the internal model and bases action selection (at the first-stage state) on both the previous outcome and the previous transition, leading to more flexible and successful behavior. Besides the distinct demands on (model-based) learning in the Markov decision task, different task designs could focus on different aspects of the internal model and necessitate the implementation of different (pre)assigned contingencies. For example, the manipulation of learnability (i.e., if contingencies between actions and outcomes are predictable or happen randomly, see Study 1 and 2), could also be incorporated in an internal model, possibly resulting in action selection that is insensitive to previous outcomes under random contingencies but sensitive to previous outcomes under predictable contingencies.

Despite the initial assumption that model-free and model-based learning are computationally separable and act on distinct (neural) representations of the task (Daw et al., 2005; Keramati, Dezfouli, & Piray, 2011), recent literature highlights the interaction between both systems to facilitate learning and solve the credit assignment problem (Moran, Keramati, Dayan, & Dolan, 2019). Moreover, recent simulation studies revealed that model-free and model-based behavior are difficult to separate in the Markov decision task under certain circumstances (Akam, Costa, & Dayan, 2015; Kool, Cushman, & Gershman, 2016). The picture is further complicated by the postulate that the representations on which both model-free and model-based reinforcement learning operate to estimate the task contingencies are themselves subject to ongoing learning and optimization (Gershman & Niv, 2010; Gershman, Norman, & Niv, 2015; Niv, 2019).

Outline of subsequent studies

Before drawing the outline of the three studies in this thesis, an explicit working definition of the central concepts under investigation is in order. *Reinforcement learning*, as defined formally and formerly (see Equations 1, 2, 3), constitutes the main framework for understanding human decision making under the premise of reward maximization. A *task representation* defines which elements (i.e., stimuli, actions, and outcomes) are relevant within a task and connects them within the central computations of reinforcement learning to allow learning of task contingencies (via the estimation of action values). If no task representation is available, *credit assignment* serves to infer a (plausible) task representation by assigning outcomes to actions. While task representations are necessary for both model-free and model-based reinforcement learning, only model-based reinforcement learning utilizes an *internal model* that (pre)specifies certain contingencies (e.g. action values, state transition probabilities) based on prior explicit knowledge from instruction or observation.

The following three studies aim to further elucidate the interaction of reinforcement learning and its task representation with internal models and credit assignment. I start my investigation with the question on how internal models can affect the task representation on which reinforcement learning operates. More specifically, I investigate how explicit knowledge about contingencies in the environment modulates the central computations of reinforcement learning, i.e., the calculation of prediction errors and the updating of action values. Subsequently, I extend this investigation to account for more complex environments which necessitate the application of more complex representations. Again, I ask how internal models influence reinforcement learning. Finally, I take a different approach and ask, how credit assignment can shape the task representation. Here, my goal is to provide insights into the possible mechanisms of credit assignment and inference which enable the emergence of appropriate task representations within a reinforcement learning perspective.

To investigate these issues, I utilize a computational modelling approach which provides both qualitative and quantitative predictions for the central computations in the brain and can thus help us understand its working principles from a mechanistic viewpoint. As I am interested in the implementation of reinforcement learning processes and concepts in the brain, I utilized a second approach and recorded electrophysiological data in addition to the behavioral data in human participants. While the behavioral data can already tell us a lot about the computational mechanisms of the brain, electrophysiological data add a further layer of insight. Especially, the integration and synthesis of these approaches, that is computational modelling, experimental manipulation, and the collection of electrophysiological data, can lead to new insights which would not be possible by each single-method approach alone. Searching for the predicted patterns of a computational model in both the behavioral and neural data of human participants is a promising means for answering questions on the realization of different computational systems under varying task conditions in the human brain.

Study 1 employs a standard bandit task in which outcomes (win or loss) are probabilistically mapped to actions. This means that the same action is not always rewarded or punished but only sometimes, based on a specific win probability. To guarantee constant learning, this probability is set to change over the course of a block according to a random walk. The main manipulation in the study affected a central characteristic of the contingency between actions and outcomes, the so-called *learnability*. In a learning condition, participants were faced with a probabilistic and volatile, yet predictable structure which was regarded as learnable. In a gambling condition, the same participants were faced with a random structure, which was regarded as unlearnable, due to the unpredictability of the action-outcome contingency. Participants were explicitly instructed on the nature of the task and the identity of the different conditions. A computational model of reinforcement learning was fit to extract latent variable estimates which were subsequently regressed on the EEG data. This approach

allows not only to find neural correlates of reinforcement learning but also to contrast the experimental conditions regarding their underlying task representations. Centrally, I investigated which subprocesses of reinforcement learning (calculation of reward prediction error or the updating of action values, equation 1 and 2 respectively) are affected by the learnability of the task. I found that the learnability of the task specifically modulates only the updating of action values (equation 2), which is interpreted as a suppression of reinforcement learning when the internal model suggests unpredictable contingencies and thus learning is not adaptive for the agent.

Following this first attempt to investigate the influence of different internal models on reinforcement learning, Study 2 extended this idea to include more complex forms of task representations as well. In order to reliably elicit such complex task representations, we adapted the Markov decision paradigm (Daw et al., 2011). Again, the experimental manipulation included a learnable condition which was contrasted by a random condition. In contrast to Study 1, participants were instructed on the structure of the task but not on the identity of the conditions, which had to be inferred from experience. As with Study 1, a computational model of reinforcement learning was fit, and latent variable estimates extracted to be used for a subsequent regression with the EEG data. Again, this approach allows not only to find neural correlates of reinforcement learning but also to contrast task representations between experimental conditions. I find that, as in Study 1, the predictability of outcomes from action in complex tasks distinctly modulates the computational processes of reinforcement learning, lending support to the idea of a strong influence of internal models on task representations and reinforcement learning.

In Study 3, I follow up on my previous research and ask how a plausible representation is selected to drive decision making when no prior knowledge about the task structure is available. In line with the idea of competitive interaction between multiple reinforcement learning systems, I sketch an inference mechanism which arbitrates between

different competing representation. To allow inference and credit assignment, the paradigm chosen for this study differs from the previous two studies and no random condition was included. However, two separate standard bandit tasks with each one transition were employed, so that on every trial, participants executed two actions and received two outcomes. Each bandit task (i.e., choice) was associated with a color-coded feedback but crucially participants were not instructed on the contingency between tasks and colors. Due to this initial uncertainty about the representation of the task, an inference mechanism is necessary to select and arbitrate control towards the most plausible representation. Again, central latent variable estimates were derived and used to predict behavioral and neural data. I find that multiple reinforcement learning systems are realized in the brain and that their respective computations of prediction errors can be used for credit assignment and the selection of the correct representation of the task. Taken together, this suggests a bidirectional interaction between reinforcement learning and task representation, which allows credit assignment under uncertainty about the correct representation of the environment.

Study 1: Task learnability modulates value updating but not prediction errors in probabilistic choice tasks

By Franz Wurm, Wioleta Walentowska, Benjamin Ernst, Mario Carlo Severo, Gilles Pourtois, and Marco Steinhauser

Abstract

The goal of reinforcement learning is to maximize outcomes and improve future decision making. In gambling tasks, however, decision making cannot be improved due to the lack of learnability. Based on the idea that reinforcement learning comprises two subprocesses (calculation of reward prediction errors and updating of action values), we asked which of these subprocesses is affected when a task is not learnable. We contrasted behavioral data and event-related potentials (ERPs) in a learning variant and a gambling variant of a simple two-armed bandit task in which outcome sequences were matched across tasks. Participants were explicitly informed that feedback could be used to improve performance in the learning task but not in the gambling task, and we predicted a corresponding modulation of the subprocesses of reinforcement learning. Based on a computational model of the two task variants, we used a model-based analysis of ERP data to extract the neural footprints of these subprocesses in the two tasks. Our results revealed that task learnability modulates reinforcement learning via the suppression of action value updating but leaves the calculation of reward prediction errors unaffected. Based on our model and the data, we propose that task learnability influences the strength of action value updating as well as the trade-off between choice policies (reinforcement learning, stochastic choice) based on a flexible cost-benefit arbitration.

This paper is currently under review at *Journal of Cognitive Neuroscience*

Study 2: The influence of internal models on feedback-related brain activity

By Franz Wurm, Benjamin Ernst, and Marco Steinhauser

Abstract

Decision making relies on the interplay between two distinct learning mechanisms, namely habitual model-free learning and goal-directed model-based learning. Recent literature suggests that this interplay is significantly shaped by the environmental structure as represented by an internal model. We employed a modified two-stage but one-decision Markov decision task to investigate how two internal models differing in the predictability of stage transitions influence the neural correlates of feedback processing. Our results demonstrate that fronto-central theta and the feedback-related negativity (FRN), two correlates of reward prediction errors in the medial frontal cortex, are independent of the internal representations of the environmental structure. In contrast, centro-parietal delta and the P3, two correlates possibly reflecting feedback evaluation in working memory, were highly susceptible to the underlying internal model. Model-based analyses of single-trial activity showed a comparable pattern, indicating that while the computation of unsigned reward prediction errors is represented by theta and the FRN irrespective of the internal models, the P3 adapts to the internal representation of an environment. Our findings further substantiate the assumption, that the feedback-locked components under investigation reflect distinct mechanisms of feedback processing and that different internal models selectively influence these mechanisms.

Published as Wurm, F., Ernst, B., & Steinhauser, M. (2020). The influence of internal models on feedback-related brain activity. *Cognitive, Affective, & Behavioral Neuroscience*, *20*, 1070-1089.

Study 3: Surprise-minimization as a solution to the structural credit assignment problem

By Franz Wurm, Benjamin Ernst, and Marco Steinhauser

Abstract

The structural credit assignment problem arises when the causal structure between actions and subsequent outcomes is hidden from direct observation. To solve this problem and enable goal-directed behavior, an agent has to infer structure and form a representation thereof. In the scope of this study, we investigate a possible solution in the human brain. We recorded behavioral and electrophysiological data from human participants in a novel variant of the bandit task, where multiple actions lead to multiple outcomes. Crucially, the mapping between actions and outcomes was hidden and not instructed to the participants. Human choice behavior revealed clear hallmarks of credit assignment and learning. Moreover, a computational model which formalizes action selection as the competition between multiple representations of the hidden structure was fit to account for participants data. Starting in a state of uncertainty about the correct representation, the central mechanism of this model is the arbitration of action control towards the representation which minimizes surprise about outcomes. Crucially, single-trial latent-variable analysis reveals that the neural patterns clearly support central quantitative predictions of this surprise minimization model. The results suggest that posterior activity is not only related to reinforcement learning under correct as well as incorrect task representations but also reflects central mechanisms of credit assignment and representation learning.

This paper is currently in preparation.

General discussion

The goal of this thesis was to elucidate the interactive nature of reinforcement learning, its task representations, internal models, and credit assignment across distinct levels of complexity. In Study 1, participants worked through learnable and random conditions of a simple bandit task. As a central result of the study, reinforcement learning was distinctly modulated between these conditions. While neural activity in the learning condition reflected both prediction errors and action values, the neural activity in the random condition reflected only prediction errors, suggesting that reinforcement learning was suppressed in this condition. Interestingly, both patterns of neural activity can be interpreted as highly adaptive within the respective task conditions and consistently reflect the impact of the internal model on reinforcement learning and its task representation. In Study 2, a comparable pattern was observed. Participants worked through learnable and random conditions of a Markov decision task, a multi-stage variant of the bandit task, which is hypothesized to elicit the applications of multiple reinforcement learning systems. Again, neural activity related to reinforcement learning showed a distinct pattern between conditions. Crucially, this effect can be again interpreted as the consequence of the internal model on reinforcement learning and its task representations. In Study 3, I investigated the reversed effect, namely how reinforcement learning can shape the task representation via credit assignment. Participants worked through a novel variant of the bandit task in which multiple decisions led to multiple outcomes, but the link between decision and outcomes was not known to the participants and thus had to be inferred from experience. As a central finding of this study, participants correctly inferred the correct mapping between decision and outcomes, in accordance with a computational model that utilized the existence of multiple reinforcement learning systems to solve the credit assignment problem and inform the representation of the task. Crucially, model predictions of

latent variables involved in this inference and credit assignment process were reflected in the neural data, showing the influence of reinforcement learning on the task representation.

The ubiquity of prediction errors

A stable finding across all three studies presented in this thesis is that prediction errors are calculated in the brain irrespective of whether these prediction errors were translated into learning. In Study 1 and 2 we replicated the common finding that the FRN reliably reflects a reward prediction error. While this finding is expected for learnable conditions, it is surprising in the random condition, where learning is impossible and a representation without (reinforcement) learning was pursued by participants. However, this finding of a dissociation between behavioral and neural patterns fits with recent research on both animals and humans. Monkeys' neural activity continued to estimate (model-free) reward prediction errors even when behavior followed a completely different (model-based) policy (Bayer & Glimcher, 2005). In humans, FRN amplitudes reflected reward prediction errors in a reversal learning task, in which task contingencies were reversed at random points, and continued to reflect prediction errors based on old action values, even when the behavior already indicated that participants assumed a contingency reversal (Chase et al., 2011). In another study, participants were instructed on the exact action values and although this internal model led to asymptotic optimal behavior, the FRN was modulated as if the task representation did not incorporate this information (Walsh & Anderson, 2011). In Study 3, I extended existing evidence on the calculation of multiple prediction errors. To my knowledge, this is the first study which explicitly tested for the existence of alternative prediction errors, which are not calculated on a correct representation of the task but follow an incorrect representation of the task. Arguably, the main reason for the calculation of prediction errors for multiple representations is the uncertainty about the correct representation. Based on the computational

model, this incorrect prediction errors (and also the correct prediction errors) provided a teaching signal which was not only used for the estimation of contingencies within the representation but also the arbitration between task representations. Thus, prediction errors can be used for multiple purposes, further substantiating their ubiquity in the human brain.

There also exist alternative instantiations of prediction errors. As in the three studies of this thesis, the most commonly employed prediction errors are *reward prediction errors*, implemented by a temporal difference process. As formalized above (see Equation 1), reward prediction errors quantify the difference between an expected value of an action and the observed outcome. While this calculation is mainly realized by model-free reinforcement learning, the model-based counterpart is called *state-prediction error* (Gläscher, Daw, Dayan, & O'Doherty, 2010): It measures the surprise of a new state given there is no external (i.e., observable) rewarding outcome (Walsh & Anderson, 2010). Recent work further suggests the existence of distinct model-based reward prediction errors at the level of outcome presentation (Sambrook et al., 2018), as well as so-called *risk-prediction errors* (Preuschoff, Bossaerts, & Quartz, 2006; Preuschoff, Quartz, & Bossaerts, 2008) which quantify the uncertainty about a certain reward prediction error. *Pseudo-reward prediction errors* (Botvinick, 2012; Botvinick, Niv, & Barto, 2009; Sutton, Precup, & Singh, 1999) are basically calculated on internal reward signals, for example after the achievement of a task subgoal. As an extension of the reinforcement learning framework, hierarchical reinforcement learning postulates the existence of such pseudo-reward prediction errors, that allow to extend learning of simple actions to account for elaborate action sequences (i.e. options). Evidence for the existence of pseudo-reward prediction errors have been found in the human brain (Diuk, Tsai, Wallis, Botvinick, & Niv, 2013; Ribas-Fernandes et al., 2011). Finally, in an overarching endeavor, the postulate of *generalized prediction errors,* which incorporate from both reward as well as sensory inputs, complements the arsenal of prediction error and calls for a reappraisal of the role of midbrain dopamine towards signaling a broader concept of

prediction error that is embedded between perception and reward processing (Gardner, Schoenbaum, & Gershman, 2018; Langdon, Sharpe, Schoenbaum, & Niv, 2018).

In sum, the three studies of this thesis concurrently line up with previous findings on the ubiquity of prediction errors in the human brain. However, especially the results from Study 3 adds nuanced evidence to this issue by showing that even prediction errors calculated from an incorrect representation of the environment are reflected in the brain, further substantiating the idea of multiple independent control systems or policies (Daw & O'Doherty, 2014).

Message passing in the brain

While there seem to be a wide variety of prediction errors in the brain, ostensibly reflected in (phasic) dopamine activity, the studies presented in this thesis also highlight the propagation and subsequent processing of prediction errors within the human brain. In line with the idea of a (hierarchical) self-supervised system (Dayan et al., 1995), my studies show the importance of both bottom-up and top-down message passing to allow goal-directed behavior. In Study 1, where participants were explicitly instructed on the learnability of the task, the utilization of this knowledge was presumably implemented by top-down connections, supposedly via pathways from the prefrontal cortex to the basal ganglia (Doll, Hutchison, & Frank, 2011; Doll, Jacobs, Sanfey, & Frank, 2009). Critically, there was an interaction between different message passing mechanisms, where putative top-down control from prefrontal areas (via the internal model) mediated the bottom-up passing of reward prediction error from midbrain areas (via the task representation and reinforcement learning). These top-down effects are likely to have led to the pronounced modulation of action value updating (Equation 2) that was observed in all three studies of this thesis. In Study 1, the updating of action values was enabled in the learnable condition, but inhibited in the gambling

condition, even though reward prediction errors were evident in both conditions. In Study 2, we found a similar effect towards a pronounced modulation only for the predictable structure, but not the random structure. In Study 3, the same effect was observed, and the updating of action values was inhibited only for the implausible but not the plausible representation of the causal structure.

A possible mechanism with a central role in such a message passing system of reinforcement learning is attention. Attention has been suggested to interact with learning in a bidirectional way (Leong, Radulescu, Daniel, DeWoskin, & Niv, 2017; Radulescu, Niv, & Ballard, 2019). On the one hand, learning can guide attention, so that dimensions that are predictive of reward are attended more strongly (Mackintosh, 1975). On the other hand, selective attention can guide learning towards specific dimensions of the environment, reducing the number of states and actions that have to be learned, hence simplifying the computations and making reinforcement learning and credit assignment more efficient (Niv et al., 2015). An additional aspect of attention which seems to be relevant for the interpretation of our results is the proposal of a distinction between attention at action and attention at feedback (Dayan, Kakade, & Montague, 2000). To maximize reward, attention during action selection should be directed towards the stimulus (feature) with the highest predictive value of reward. To minimize uncertainty, attention during feedback processing should be directed towards the most surprising outcome (features). Based on this idea that attention should be guided by two separate goals (i.e., optimize action selection and maximize information), the distinct subprocesses of reinforcement learning could be directly mapped onto the different roles of attention: While the calculation of prediction errors might be associated with the attention at feedback in order to extract information, updating of action values might be associated with the attention at action in order to optimize action selection. Arguably, the manipulation of learnability in Study 1 and 2 could have also impacted the attention at action. For example, in Study 1, participants were explicitly instructed that they “cannot in any way

influence the outcome". Thus, if the distinct roles of attentions hold to be true, this might explain the intricate pattern of the electrophysiological data.

Besides the role of attention, a hierarchical architecture is a natural extension of the computational model of reinforcement learning, and hierarchical tasks and models have already significantly improved our understanding of neural networks and message passing in the brain (Botvinick, 2012; Koechlin & Summerfield, 2007; Pezzulo, Rigoli, & Friston, 2018). In addition to the previous discussion of the framework of hierarchical reinforcement learning (Botvinick, 2012; Botvinick et al., 2009), the importance of hierarchical organization is further supported by various examples across (cognitive) neuroscience (Friston, 2008, 2010), where the architecture of cortical sensory areas strongly supports the notion of hierarchy, in which messages are passed and integrated between functionally segregated areas to produce complex behavior (Zeki & Shipp, 1988) and hierarchical principles are commonly used to enhance the statistical power or allow the estimation of individual subject parameters as well as group distributions (Gelman, 2008; Kruschke, 2014; Vandekerckhove, Tuerlinckx, & Lee, 2011; Wagenmakers & Lee, 2013; Wiecki, Sofer, & Frank, 2013).

Constructing better models

In the scope of this thesis, I explored a selected subset of possible architectures and algorithmic processes for solving complex decision and inference problems. The core mechanisms relied on two simple processes commonly used in the reinforcement learning framework. First, the temporal difference rule allowed to estimate and update the expected value of an action as describer earlier (Equation 1 and 2). Second, the softmax rule (Equation 3) translated these estimated values into action probabilities, allowing a policy to constantly optimize its associated behavior. While my fine tailored solutions for the specific experimental tasks were well suited to explain behavioral differences and track latent

variables at the neural level, there exist infinite further possible realizations of processes, all of which could be tested to improve the fit between computational models and the data. An advantage of the model-based approach is that every possible instantiation could separately induce knowledge about the algorithmic workings of the brain (Nassar & Frank, 2016). Of course, this inductive process should always be balanced by critical deduction and a falsification approach (Popper, 2005), to prevent spurious interpretation of models and estimated parameters (Nassar & Frank, 2016). Based on the models within this thesis I will highlight three important ways to improve existing architectures.

First, one can identify central processes of different models and completely switch them with other possible processes. Take, for example, the softmax process, which transforms action values into action probabilities. Even within reinforcement learning there exist multiple different candidate processes for action selection. One prominent alternative is the so-called greedy action selection (Sutton & Barto, 2018). Greedy selection always selects the action with the highest value for execution. Although this selection process exploits current knowledge about action values, it completely neglects exploring alternative actions. In contrast to fully exploitative behavior, one could also construct a fully random and explorative choice strategy, as was done in Study 1 and 2. Interestingly, the softmax selection can mimic both greedy action selection and fully random behavior by setting the inverse temperature parameter either high or close to zero. An alternative for action selection which cannot be subsumed by softmax selection is drift diffusion modelling (DDM, Ratcliff, 1978). DDM is a widely used sequential-sampling model (Cavanagh et al., 2011; Forstmann, Ratcliff, & Wagenmakers, 2016; Ratcliff & McKoon, 2008; Wabersich & Vandekerckhove, 2014), which assumes that action selection is determined by continuously sampling noisy evidence until a decision boundary is reached in favor of one action. Although DDM needs up to 4 free parameters (while softmax only needs one), its main advantage is the extraction of information about decision making not only from accuracy but also from response time data.

Because response times have so far not received much attention in the reinforcement learning literature (Keramati et al., 2011), the implementation of DDM can yield new insights into the neural underpinnings of reinforcement learning (Frank et al., 2015) as well as psychopathology (Pedersen, Frank, & Biele, 2016). Moreover, a recent study suggests that the use of DDM substantially improves the parameter recovery and stability of individual estimates of the trade-off between model-free and model-based learning for human participants in the Markov decision paradigm (Shahar et al., 2019).

Second, computational models should be carefully reconciled with the task structure. Computational models usually consist of different free parameters, which are estimated from either behavioral or neural data. These parameters then drive central processes that relevantly contribute to the model's characteristic pattern of decision making in the task. For example, within the temporal difference process, the learning rate controls the updating of the action values by the most recent prediction error (see Equation 2). Therefore, the learning rate constitutes a central parameter within my experiments. Crucially, this free model parameter was estimated as a constant scalar, fixed across the whole experiment for each participant. Although this is a common approach in reinforcement learning to induce an explanation for individual or condition-dependent differences (e.g. M. X. Cohen, 2007; Otto, Gershman, et al., 2013), a recent study suggests that the interpretation of overly simplified models (e.g. with a fixed learning rate) can be misleading as it biases the estimation of model parameters (Nassar & Gold, 2013). Using an estimation task in which participants had to predict upcoming reward magnitude, the authors demonstrated that although fixed learning rates have a good account for behavior, only variable learning rates can account for the increase of behavioral adaptation following sudden changes in action-outcome contingency (Jepma et al., 2016; Nassar et al., 2019; Nassar & Gold, 2013). However, before coming to rash conclusions about the validity of fixed learning rates in reinforcement learning, one should pay attention to the importance of task design. In contrast to the studies presented in this thesis, in which

reward contingencies were either fully fixed or constantly changing, the contingencies of the estimation task were subject to random change points. In such an environment, a variable learning rate which adapts to the surprise associated with observed outcomes (Pearce & Hall, 1980) is highly plausible as it balances the need for rapid learning after rare change points in an otherwise stable environment. Based on this obvious interaction between task design and requirements for the decision maker, computational modelling should not replace thorough experimental manipulation but rather complement it to isolate the processes of interest (Nassar & Gold, 2013; Wilson & Collins, 2019).

So far, the two approaches for extending and improving computational models maintained the essential mechanisms (i.e., temporal difference and softmax) of reinforcement learning and either altered distinct processes within the model to better account for empirical data or constructed meaningful tasks to investigate the processes of interest. However, there exist multiple models of decision making that are distinct to reinforcement learning. For example, a large body of alternative models seeks to explain human probability matching. Probability matching describes the observation that participants match the action probabilities with the reward probabilities (B. R. Newell, Koehler, James, Rakow, & van Ravenzwaaij, 2013; Otto, Taylor, & Markman, 2011; Vulkan, 2000). It has been shown that matching behavior can arise under different models, such as win-stay lose-shift (WSLS: Herrnstein, 2000) or expectation matching (Sugrue, Corrado, & Newsome, 2004). Taking WSLS literally, it assumes that participants switch actions after losses but stick with actions after wins. Under expectation matching, the agent is hypothesized to integrate a moving window of past rewards, on which action selection is then based. Although decision making under these models was found to mirror probability matching in empirical data, it has been suggested that reinforcement learning best accounts for a variety of findings (Feher Da Silva, Victorino, Caticha, & Baldo, 2017) or is at least partly responsible for behavior (Worthy & Maddox, 2014). Interestingly for the scope of this thesis, it has been suggested that the main reason for

probability matching is uncertainty about the generative process of the task which thus leads to pattern search, exploration, and recency effects (Feher Da Silva et al., 2017).

Taken together, this short overview on the diverse landscape of computational modelling illustrates how existing models can be extended to further improve our understanding of decision making and cognition in the brain. Future model-based research should embrace both the advantages and pitfalls of the approach when designing experiments and interpreting results in terms of cognitive processes (Mars, Shea, Kolling, & Rushworth, 2012; Wilson & Niv, 2015). Please note that these suggestions by no means diminish their conclusiveness if model-based analyses are combined with sound experimental manipulation (Nassar & Gold, 2013). Often, even gross errors in parameter estimates result only in comparably insignificant changes in the connection between neural activity and latent variables (Wilson & Niv, 2015).

Accounting for uncertainty

A major point for criticism of all the computational models of reinforcement learning reported so far is that they only consider estimates of prediction errors or action values as simple scalars. Although the temporal difference model has proven to be a reliable generalization of earlier (reinforcement) learning models (e.g., the Rescorla-Wagner model) and is grounded in the normative theory of reinforcement learning (Gershman, 2015), it only encompasses so-called point estimator statistics (e.g., mean or variance). However, there is reliable evidence that, besides the estimation of the single most likely values for central latent variables, the brain also incorporates the uncertainty or precision about these estimates (Bach & Dolan, 2012; Pouget, Beck, Ma, & Latham, 2013). As we have seen, uncertainty about contingency (Studies 1-3) and especially structure (Study 3) is an important component in efficient reinforcement and therefore deserves further considerations. A more fine-grained

distinction dichotomized expected and unexpected uncertainty (Yu & Dayan, 2005). Under this notion, reward prediction errors might be simple scalar values to update value functions, but the system additionally utilized these prediction errors to estimate uncertainty.

One possible mechanism proposes a novel set of latent variables which is incorporated in the temporal difference framework: outcome variance and risk prediction errors (Preuschoff et al., 2006, 2008). Crucially, these variables work similarly to the already established latent variables of action values and reward prediction errors. The risk prediction error is calculated as the squared prediction error normalized by the mean outcome and therefore is closely linked to surprise. Outcome variance is then estimated by incrementally updating with the risk prediction errors. For the agent, the estimated outcome variance constitutes a good measure of expected uncertainty. Unexpected uncertainty can then be regarded as the surplus of the trial-to-trial risk prediction error that is unexplained by the estimated outcome variance or expected uncertainty. Foremost, high unexpected uncertainty indicates an inadequate representation of the task's structure. In line with the finding from Study 3 that the competition between (absolute) prediction errors is reflected in the brain, the demonstration that risk prediction errors are reflected in neural structures such as the anterior insula (Preuschoff et al., 2006, 2008) or ventral striatum (d'Acremont, Lu, Li, Van der Linden, & Bechara, 2009) further consolidate the idea that the brain implements reinforcement learning as a point-estimation of value, with a limited estimation of precision (Findling, Skvortsova, Dromnelle, Palminteri, & Wyart, 2019)

Another possible mechanism which naturally incorporates the notion of uncertainty is Bayesian inference. In contrast to reinforcement learning models in which the ultimate goal is to optimize the long-term reward within a task, Bayesian models foremost deal with inference about structured knowledge. So far, their use was mainly limited to the domain of category learning (Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Tenenbaum, Griffiths, & Kemp, 2006). In a category learning task, participants must ascribe stimuli to different

categories based on the presence or absence of a specific stimulus feature. Similar to inference in Study 3, the initial uncertainty about category membership or the defining stimulus feature must be disambiguated. Although Bayesian models still lack a plausible neurobiological implementation (Gershman, 2015; Radulescu et al., 2019), their use within neuroscientific research of decision making is promoted by findings that humans act in accordance with Bayesian optimality. Because a full Bayesian approach becomes intractable with complexity (Kwisthout, Wareham, & van Rooij, 2011), the usage of approximation methods in neuroscience is increasingly popular (Sanborn & Chater, 2016; Sanborn, Griffiths, & Navarro, 2010). One variant are particle filters (Gershman, 2015; Radulescu et al., 2019) or variational Bayes (Friston et al., 2015; Mathys, 2011; Sajid, Ball, & Friston, 2020). These methods act in accordance with Bayes theorem, by transforming a prior distribution over the belief about different candidate representations into a posterior distribution by updating with a likelihood that states the probability of the observations given the representation (the filter).

While reinforcement learning and Bayes inference have long been considered in independent domains, there is an ongoing urge to integrate both computational approaches (Gershman, 2015; Radulescu et al., 2019). On the one hand, reinforcement learning is an efficient method to optimize decision making and minimize expected uncertainty under a specific representation or internal model. On the other hand, Bayesian inference is a potent tool to infer and arbitrate between representations and adapt to unexpected uncertainty. The central idea behind a unification of both approaches is that the internal models are learned through (approximate) Bayesian inference and subsequently used as a source of top-down modulation that shape the representation over which reinforcement learning is optimizing decision making.

Conclusion

In the scope of this thesis, I used computational modelling to formalize and capture the interactions between reinforcement learning and the representation of the environment across increasingly complex task environments. Crucially, the computational modelling approach allowed me to derive latent variables from the behavioral data which were subsequently used to reveal patterns in the neural data that reflected central model computations. In three EEG-studies I showed that the quantitative predictions from the computational model of reinforcement learning were evident on the neural level. Crucially, goal-directed modulations within the neural computations indicated an elaborate interplay of cooperation and competition between separable processes and systems of reinforcement learning. In Study 1 and 2, the internal model was identified as the driving neural modulator of reinforcement learning and the top-down implementation of control. A central finding was the biasing of the (ubiquitous) temporal difference process, which can be interpreted as a necessary step for optimizing reward and uncertainty across different levels of task complexity. In Study 3, a modified design of the classic bandit task, in which multiple actions lead to multiple outcomes and the correct representation of the mapping between actions and outcomes are unknown, allowed us to implement a novel computational architecture which uses reinforcement learning principles to infer credit assignment and inform the task representation. Critically, this novel inference model makes distinct predictions for higher-level computations and credit assignment, expressed in an uncertainty formulation. The predicted surprise-based computations were clearly evident in posterior neural data. As the electrode sites for the surprise-based inference signal match those of both earlier studies of this thesis and the literature (Jepma et al., 2018, 2016; Nassar et al., 2019; Polich, 2007), this further adds to the importance of this electrophysiological phenomenon in the study of the brain as an information processing system (Polich, 2020).

Taken together, this thesis provides a step towards an integration of multiple functional principles of the brain in a biologically plausible framework. Besides the importance of appropriate task designs, this empirical work puts an emphasis on the indispensable role of computational models to disentangle the intricate cortical operations thus allowing us to study the brain under a holistic perspective (Radulescu et al., 2019; Bogacz, 2017; Friston, Daunizeau, & Kiebel, 2009; Sajid, Ball, & Friston, 2020). Reinforcement learning, and more specifically the temporal difference algorithm, has successfully paved the way to a biologically plausible description of the human brain as an extremely powerful and efficient information processing system. However, more research is required to further our understanding how the brain can actively control its internal computations and representations in a goal-directed manner.

References

Adams, C. D. (1982). Variations in the sensitivity of instrumental responding to reinforcer devaluation. *The Quarterly Journal of Experimental Psychology Section B*, *34*, 77–98.

Adams, C. D., & Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *The Quarterly Journal of Experimental Psychology Section B*, *33*, 109–121.

Akam, T., Costa, R., & Dayan, P. (2015). Simple plans or sophisticated habits? State, transition and learning interactions in the two-step task. *PLoS Computational Biology*, *11*, 1–25.

Asaad, W. F., Lauro, P. M., Perge, J. A., & Eskandar, E. N. (2017). Prefrontal neurons encode a solution to the credit-assignment problem. *The Journal of Neuroscience*, *37*, 6995–7007.

Bach, D. R., & Dolan, R. J. (2012). Knowing how much you don't know: A neural organization of uncertainty estimates. *Nature Reviews Neuroscience*, *13*, 572–586.

Balleine, B. W. (2005). Neural bases of food-seeking: Affect, arousal and reward in corticostriatolimbic circuits. *Physiology and Behavior*, *86*, 717–730.

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37*, 407–419.

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, *47*, 129–141.

Bayer, H. M., Lau, B., & Glimcher, P. W. (2007). Statistics of midbrain dopamine neuron spike trains in the awake primate. *Journal of Neurophysiology*, *98*, 1428–1439.

Bellman, R. (1957). Functional equations in the theory of dynamic programming--VII. A partial differential equation for the Fredholm resolvent. *Proceedings of the American*

Mathematical Society, *8*, 435.

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

Berns, G. S., McClure, S. M., Pagnoni, G., & Montague, P. R. (2001). Predictability modulates human brain response to reward. *Journal of Neuroscience*, *21*, 2793–2798.

Berry, D. A., & Fristedt, B. (1986). Bandit problems: Sequential allocation of experiments, *149*, 271.

Blodgett, H. C. (1929). The effect of the introduction of reward upon the maze performance of rats. *University of California Publications in Psychology*, *4*, 113–134.

Bogacz, R. (2017). A tutorial on the free-energy framework for modelling perception and learning. *Journal of Mathematical Psychology*, *76*, 198–211.

Botvinick, M. M. (2012). Hierarchical reinforcement learning and decision making. *Current Opinion in Neurobiology*, *22*, 956–962.

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

Cavanagh, J. F. (2015). Cortical delta activity reflects reward prediction error and related behavioral adjustments, but at different times. *NeuroImage*, *110*, 205–216.

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

Chase, H. W., Swainson, R., Durham, L., Benham, L., & Cools, R. (2011). Feedback-related negativity codes prediction error but not behavioral adjustment during probabilistic

reversal learning. *Journal of Cognitive Neuroscience*, *23*, 936–946.

Cockburn, J., & Holroyd, C. B. (2018). Feedback information and the reward positivity. *International Journal of Psychophysiology*, *132*, 243–251.

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*, 933–942.

Cohen, M. X. (2007). Individual differences and the neural representations of reward expectation and reward prediction error. *Social Cognitive and Affective Neuroscience*, *2*, 20–30.

Corbit, L. H., & Balleine, B. W. (2003). The role of prelimbic cortex in instrumental conditioning. *Behavioural Brain Research*, *146*, 145–157.

d'Acremont, M., Lu, Z. L., Li, X., Van der Linden, M., & Bechara, A. (2009). Neural correlates of risk prediction error during reinforcement learning in humans. *NeuroImage*, *47*, 1929–1939.

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

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

Daw, N., & O'Doherty, J. P. (2014). Multiple systems for value learning. In *Neuroeconomics* (pp. 393–410). Elsevier.

Daw, N., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*, 876–879.

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The Helmholtz machine.

Neural Computation, *7*, 889–904.

Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention. *Nature Neuroscience*, *3*, 1218–1223.

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology*, *18*, 185–196.

Dickinson, A., & Balleine, B. W. (1995). Motivational control of instrumental action. *Current Directions in Psychological Science*, *4*, 162–167.

Dickinson, A., & Balleine, B. W. (2002). The role of learning in the operation of motivational systems. In C. R. Gallistel (Ed.), *Steven's handbook of experimental psychology: Learning, motivation and emotion* (Vol. 3, pp. 497–534). New York: Wiley.

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

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*, 312–325.

Doll, B. B., Hutchison, K. E., & Frank, M. J. (2011). Dopaminergic genes predict individual differences in susceptibility to confirmation bias. *Journal of Neuroscience*, *31*, 6188–6198.

Doll, B. B., Jacobs, W. J., Sanfey, A. G., & Frank, M. J. (2009). Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. *Brain Research*, *1299*, 74–79.

Duncan-Johnson, C. C., & Donchin, E. (1977). On quantifying surprise: The variation of event-related potentials with subjective probability. *Psychophysiology*, *14*, 456–467.

Feher Da Silva, C., Victorino, C. G., Caticha, N., & Baldo, M. V. C. (2017). Exploration and

recency as the main proximate causes of probability matching: A reinforcement learning analysis. *Scientific Reports*, *7*, 1–23.

Findling, C., Skvortsova, V., Dromnelle, R., Palminteri, S., & Wyart, V. (2019). Computational noise in reward-guided learning drives behavioral variability in volatile environments. *Nature Neuroscience*, *22*, 2066–2077.

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward dopamine neurons. *Science*, *299*, 1898–1902.

Fischer, A. G., & Ullsperger, M. (2013). Real and fictive outcomes are processed differently but converge on a common adaptive mechanism. *Neuron*, *79*, 1243–1255.

FitzGerald, T. H. B., Seymour, B., & Dolan, R. J. (2009). The role of human orbitofrontal cortex in value comparison for incommensurable objects. *Journal of Neuroscience*, *29*, 8388–8395.

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual Review of Psychology*, *67*, 641–666.

Frank, M. J., Gagne, C., Nyhus, E., Masters, S., Wiecki, T. V., Cavanagh, J. F., & Badre, D. (2015). FMRI and EEG predictors of dynamic decision parameters during human reinforcement learning. *Journal of Neuroscience*, *35*, 485–494.

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By carrot or by stick: Cognitive reinforcement learning in Parkinsonism. *Science*, *306*, 1940–1943.

Friston, K. J. (2008). Hierarchical models in the brain. *PLoS Computational Biology*, *4*.

Friston, K. J. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*, 127–138.

Friston, K. J., Daunizeau, J., & Kiebel, S. J. (2009). Reinforcement learning or active inference? *PLoS ONE*, *4*.

Friston, K. J., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, *6*, 187–214.

Gardner, M. P. H., Schoenbaum, G., & Gershman, S. J. (2018). Rethinking dopamine as generalized prediction error. *Proceedings of the Royal Society B: Biological Sciences*, *285*, 20181645.

Gazzaniga, M., Ivry, R., & Mangun, G. (2014). *Cognitive neuroscience: The biology of the mind* (4th ed.). New York: W.W. Norton.

Gelman, A. (2008). Data analysis using regression and multilevel/hierarchical models. Cambridge: Cambridge University Press.

Gershman, S. J. (2015). A unifying probabilistic view of associative learning. *PLoS Computational Biology*, *11*, 1–20.

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: Carving nature at its joints. *Current Opinion in Neurobiology*, *20*, 251–256.

Gershman, S. J., Norman, K. A., & Niv, Y. (2015). Discovering latent causes in reinforcement learning. *Current Opinion in Behavioral Sciences*, *5*, 43–50.

Gillan, C. M., Otto, A. R., Phelps, E. A., & Daw, N. (2015). Model-based learning protects against forming habits. *Cognitive, Affective, & Behavioral Neuroscience*, *15*, 523–536.

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

Gläscher, J. P., & O'Doherty, J. P. (2010). Model-based approaches to neuroimaging:

Combining reinforcement learning theory with fMRI data. *Wiley Interdisciplinary Reviews: Cognitive Science*.

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 15647–15654.

Goodman, N. A., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*, 108–154.

Graybiel, A. M. (2008). Habits, rituals, and the evaluative Brain. *Annual Review of Neuroscience*, *31*, 359–387.

Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2006). The feedback-related negativity reflects the binary evaluation of good versus bad outcomes. *Biological Psychology*, *71*, 148–154.

Herrnstein, R. J. (2000). *The matching law: Papers in psychology and economics*. (H. Rachlin & D. Laibson, Eds.). Cambridge: Harvard University Press.

Hollerman, J. R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, *1*, 304–309.

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*, 679–709.

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

Hull, C. (1932). The goal-gradient hypothesis and maze learning. *Psychological Review*, *39*, 25–43.

Hull, C. (1942). Conditioning: Outline of a systematic theory of learning. In *The forty-first yearbook of the National Society for the Study of Education: Part II, The psychology of learning.* (pp. 61–95). Chicago: University of Chicago Press.

Jepma, M., Brown, S. B. R. E., Murphy, P. R., Koelewijn, S. C., de Vries, B., van den Maagdenberg, A. M., & Nieuwenhuis, S. (2018). Noradrenergic and cholinergic modulation of belief updating. *Journal of Cognitive Neuroscience*, *30*, 1803–1820.

Jepma, M., Murphy, P. R., Nassar, M. R., Rangel-Gomez, M., Meeter, M., & Nieuwenhuis, S. (2016). Catecholaminergic regulation of learning rate in a dynamic environment. *PLoS Computational Biology*, *12*, 1–24.

Jocham, G., Brodersen, K. H., Constantinescu, A. O., Kahn, M. C., Ianni, A. M., Walton, M. E., … Behrens, T. E. (2016). Reward-guided learning with and without causal attribution. *Neuron*, *90*, 177–190.

Johnson, A., & Redish, A. D. (2007). Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *Journal of Neuroscience*, *27*, 12176–12189.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237–285.

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Computational Biology*, *7*.

Knutson, B., Adams, C. M., Fong, G. W., & Hommer, D. (2001). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *The Journal of Neuroscience*, *21*, 1–5.

Koechlin, E., & Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences*, *11*, 229–235.

Kolossa, A., Fingscheidt, T., Wessel, K., & Kopp, B. (2012). A model-based approach to trial-by-trial P300 amplitude fluctuations. *Frontiers in Human Neuroscience*, *6*, 1–28.

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

Kool, W., Cushman, F. A., & Gershman, S. J. (2016). When does model-based control pay off? *PLoS Computational Biology*, *12*, 1–34.

Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological Science*, *28*, 1321–1333.

Kool, W., Gershman, S. J., & Cushman, F. A. (2018). Planning complexity registers as a cost in metacontrol. *Journal of Cognitive Neuroscience*, *30*, 1391–1404.

Kopp, B., Seer, C., Lange, F., Kluytmans, A., Kolossa, A., Fingscheidt, T., & Hoijtink, H. (2016). P300 amplitude variations, prior probabilities, and likelihoods: A Bayesian ERP study. *Cognitive, Affective, & Behavioral Neuroscience*, *16*, 911–928.

Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, *21*, 1148–1160.

Krigolson, O. E. (2017). Event-related brain potentials and the study of reward processing: Methodological considerations. *International Journal of Psychophysiology*, 1–9.

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan, second edition*. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan, Second Edition* (2nd ed.). London: Academic Press.

Kwisthout, J., Wareham, T., & van Rooij, I. (2011). Bayesian intractability is not an ailment that approximation can cure. *Cognitive Science*, *35*, 779–784.

Langdon, A. J., Sharpe, M. J., Schoenbaum, G., & Niv, Y. (2018). Model-based predictions

for dopamine. *Current Opinion in Neurobiology*, *49*, 1–7.

Lee, M. D., Zhang, S., Munro, M., & Steyvers, M. (2011). Psychological models of human and optimal performance in bandit problems. *Cognitive Systems Research*, *12*, 164–174.

Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron*, *93*, 451–463.

Li, J., Delgado, M. R., & Phelps, E. A. (2011). How instructed knowledge modulates the neural systems of reward learning. *Proceedings of the National Academy of Sciences*, *108*, 55–60.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*, 276–298.

Macready, W. G., & Wolpert, D. H. (1998). Bandit problems and the exploration/exploitation tradeoff. *IEEE Transactions on Evolutionary Computation*, *2*, 2–22.

Mars, R. B., Debener, S., Gladwin, T. E., Harrison, L. M., Haggard, P., Rothwell, J. C., & Bestmann, S. (2008). Trial-by-trial fluctuations in the event-related electroencephalogram reflect dynamic changes in the degree of surprise. *Journal of Neuroscience*, *28*, 12539–12545.

Mars, R. B., Shea, N. J., Kolling, N., & Rushworth, M. F. (2012). Model-based analyses: Promises, pitfalls, and example applications to the study of cognitive control. *Quarterly Journal of Experimental Psychology*, *65*, 252–267.

Mathys, C. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, *5*, 1–20.

Miller, G. A. (2003). The cognitive revolution: A historical perspective. *Trends in Cognitive*

Sciences, *7*, 141–144.

Miltner, W. H. R., Braun, C. H., & Coles, M. G. H. (1997). Event-related brain potentials following incorrect feedback in a time-estimation task: Evidence for a "generic" neural system for error detection. *Journal of Cognitive Neuroscience*, *9*, 788–798.

Minsky, M. (1961). Steps toward artificial intelligence. *Proceedings of the IRE*, *49*, 8–30.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of Neuroscience*, *16*, 1936–1947.

Moran, R., Keramati, M., Dayan, P., & Dolan, R. J. (2019). Retrospective model-based inference guides model-free credit assignment. *Nature Communications*, *10*, 750.

Nassar, M. R., Bruckner, R., & Frank, M. J. (2019). Statistical context dictates the relationship between feedback-related EEG signals and learning. *ELife*, *8*, 1–26.

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

Nassar, M. R., & Gold, J. I. (2013). A healthy fear of the unknown: Perspectives on the interpretation of parameter fits from computational models in neuroscience. *PLoS Computational Biology*, *9*, 1–6.

Newell, A., & Simon, H. A. (1972). *Human problem solving* (Vol. 104). Englewood Cliffs, NJ: Prentice-Hall.

Newell, B. R., Koehler, D. J., James, G., Rakow, T., & van Ravenzwaaij, D. (2013). Probability matching in risky choice: The interplay of feedback and strategy availability. *Memory and Cognition*, *41*, 329–338.

Nieuwenhuis, S., Holroyd, C. B., Mol, N., & Coles, M. G. H. (2004). Reinforcement-related brain potentials from medial frontal cortex: Origins and functional significance. *Neuroscience and Biobehavioral Reviews*, *28*, 441–448.

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

Niv, Y. (2019). Learning task-state representations. *Nature Neuroscience*, *22*, 1544–1553.

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

Niv, Y., & Langdon, A. J. (2016). Reinforcement learning with Marr. *Current Opinion in Behavioral Sciences*, *11*, 67–73.

Noonan, M. A., Chau, B. K., Rushworth, M. F., & Fellows, L. K. (2017). Contrasting effects of medial and lateral orbitofrontal cortex lesions on credit assignment and decision-making in humans. *The Journal of Neuroscience*, *37*, 7023–7035.

Noonan, M. A., Walton, M. E., Behrens, T. E., Sallet, J., Buckley, M. J., & Rushworth, M. F. (2010). Separate value comparison and learning mechanisms in macaque medial and lateral orbitofrontal cortex. *Proceedings of the National Academy of Sciences*, *107*, 20547–20552.

Norman, N. A. (1976). *Memory and attention: An introduction to human information processing* (2nd ed.). John Wiley & Sons.

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

O'Doherty, J. P., Hampton, A. N., & Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences*, *1104*, 35–53.

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive Map* (Vol. 27). London: Oxford University Press.

Olds, J., & Milner, P. (1954). Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of Comparative and Physiological Psychology*, *47*, 419–427.

Otto, A. R., Gershman, S. J., Markman, A. B., & Daw, N. (2013). The curse of planning: Dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological Science*, *24*, 751–761.

Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A., & Daw, N. (2013). Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences*, *110*, 20941–20946.

Otto, A. R., Taylor, E. G., & Markman, A. B. (2011). There are at least two kinds of probability matching: Evidence from a secondary task. *Cognition*, *118*, 274–279.

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

Pavlov, I. P. (1927). *Conditioned reflexes (Translated by G. V. Anrep)*. London: Oxford University Press.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*, 532–552.

Pedersen, M. L., Frank, M. J., & Biele, G. (2016). The drift diffusion model as the choice rule in reinforcement learning. *Psychonomic Bulletin & Review*.

Pezzulo, G., Rigoli, F., & Friston, K. J. (2018). Hierarchical active inference: A theory of motivated control. *Trends in Cognitive Sciences*, *22*, 294–306.

Pfeiffer, B. E., & Foster, D. J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*, *497*, 74–79.

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology*, *118*, 2128–2148.

Polich, J. (2020). 50+ years of P300: Where are we now? *Psychophysiology*, *57*, 2–3.

Popper, K. (2005). *The logic of scientific discovery* (Vol. 3). Chichester: Routledge.

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: Knowns and unknowns. *Nature Neuroscience*, *16*, 1170–1178.

Preuschoff, K., Bossaerts, P., & Quartz, S. R. (2006). Neural differentiation of expected reward and risk in human subcortical structures. *Neuron*, *51*, 381–390.

Preuschoff, K., Quartz, S. R., & Bossaerts, P. (2008). Human insula activation reflects risk prediction errors as well as risk. *Journal of Neuroscience*, *28*, 2745–2752.

Price, J. L. (2007). Definition of the orbital cortex in relation to specific connections with limbic and visceral structures and other cortical regions. *Annals of the New York Academy of Sciences*, *1121*, 54–71.

Pritchard, W. S. (1981). Psychophysiology of P300. *Psychological Bulletin*, *89*, 506–540.

Proudfit, G. H. (2015). The reward positivity: From basic research on reward to a biomarker for depression. *Psychophysiology*, *52*, 449–459.

Radulescu, A., Niv, Y., & Ballard, I. C. (2019). Holistic reinforcement learning: The role of structure and attention. *Trends in Cognitive Sciences*, *xx*, 1–15.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*, 873–922.

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

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

Roelfsema, P. R., & Holtmaat, A. (2018). Control of synaptic plasticity in deep cortical networks. *Nature Reviews Neuroscience*, *19*, 166–180.

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

Romo, R., & Schultz, W. (1990). Dopamine neurons of the monkey midbrain: Contingencies of responses to active touch during self-initiated arm movements. *Journal of Neurophysiology*, *63*, 592–606.

Sajid, N., Ball, P. J., & Friston, K. J. (2020). Active inference : demystified and compared. *ArXiv Preprint ArXiv:1909.10863*.

Sambrook, T. D., & Goslin, J. (2015). A neural reward prediction error revealed by a meta-

analysis of ERPs using great grand averages. *Psychological Bulletin*, *141*, 213–235.

Sambrook, T. D., Hardwick, B., Wills, A. J., & Goslin, J. (2018). Model-free and model-based reward prediction errors in EEG. *NeuroImage*, *178*, 162–171.

San Martín, R. (2012). Event-related potential studies of outcome processing and feedback-guided learning. *Frontiers in Human Neuroscience*, *6*, 1–17.

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*, 883–893.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*, 1144–1167.

Schultz, W. (2016). Dopamine reward prediction error coding. *Dialogues in Clinical Neuroscience*, *18*, 23–32.

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

Schultz, W., & Romo, R. (1992). Role of primate basal ganglia and frontal cortex in the internal generation of movements - I. Preparatory activity in the anterior striatum. *Experimental Brain Research*, *91*, 363–384.

Seer, C., Lange, F., Boos, M., Dengler, R., & Kopp, B. (2016). Prior probabilities modulate cortical surprise responses: A study of event-related potentials. *Brain and Cognition*, *106*, 78–89.

Shahar, N., Hauser, T. U., Moutoussis, M., Moran, R., Keramati, M., Consortium, N. S. P. N., & Dolan, R. J. (2019). Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling.

PLoS Computational Biology, *15*, 1–25.

Simon, H. A. (1978). Information-processing theory of human problem solving. In W. K. Estes (Ed.), *Handbook of Learning and Cognitive Processes* (Vol. V, pp. 271–295). Hillsdale, NJ: Lawrence Erlbaum Associates.

Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*, 168–179.

Stolyarova, A. (2018). Solving the credit assignment problem with the prefrontal cortex. *Frontiers in Neuroscience*, *12*.

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2004). Matching behavior and the representation of value in the parietal cortex. *Science*, *304*, 1782–1787.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). Cambridge: The MIT Press.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, *112*, 181–211.

Takikawa, Y., Kawagoe, R., & Hikosaka, O. (2004). A possible role of midbrain dopamine neurons in short- and long-term adaptation of saccades to position-reward mapping. *Journal of Neurophysiology*, *92*, 2520–2529.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*, 309–318.

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

Thorndike, E. L. (1911). *Animal Intelligence: Experimental Studies*. New York: The Machmillan Company.

Thorndike, E. L. (1933). A proof of the law of effect. *Science*, *77*, 173–175.

Tobler, P. N., Dickinson, A., & Schultz, W. (2003). Coding of predicted reward omission by dopamine neurons in a conditioned inhibition paradigm. *Journal of Neuroscience*, *23*, 10402–10410.

Tolman, E. C. (1948). Cognitive maps in rats and men. *The Psychological Review*, *55*, 189–208.

Turner, R. (2016). Uses, misuses, new uses and fundamental limitations of magnetic resonance imaging in cognitive science. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*.

van der Meer, M. A. A., & Redish, A. D. (2009). Low and high gamma oscillations in rat ventral striatum have distinct relationships to behavior, reward, and spiking activity on a learned spatial decision task. *Frontiers in Integrative Neuroscience*, *3*, 1–19.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, *16*, 44–62.

von Helmholtz, H. (1909). *Treatise on physiological optics* (3rd ed.). Hamburg: Voss.

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys*, *14*, 101–118.

Wabersich, D., & Vandekerckhove, J. (2014). Extending JAGS: A tutorial on adding custom distributions to JAGS (with a diffusion model example). *Behavior Research Methods*, *46*, 15–28.

Wagenmakers, E. J., & Lee, M. D. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge: Cambridge University Press.

Walsh, M. M., & Anderson, J. R. (2010). Neural correlates of temporal credit assignment.

10th International Conference on Cognitive Modeling, ICCM 2010, 265–270.

Walsh, M. M., & Anderson, J. R. (2011). Modulation of the feedback-related negativity by instruction and experience. *Proceedings of the National Academy of Sciences*, *108*, 19048–19053.

Walsh, M. M., & Anderson, J. R. (2012). Learning from experience: Event-related potential correlates of reward processing, neural adaptation, and behavioral choice. *Neuroscience and Biobehavioral Reviews*, *36*, 1870–1884.

Wassum, K. M., Cely, I. C., Maidment, N. T., & Balleine, B. W. (2009). Disruption of endogenous opioid activity during instrumental learning enhances habit acquisition. *Neuroscience*, *163*, 770–780.

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in Python. *Frontiers in Neuroinformatics*, *7*, 1–10.

Wilson, R. C., & Collins, A. G. E. (2019). Ten simple rules for the computational modeling of behavioral data. *ELife*, *8*, 1–33.

Wilson, R. C., & Niv, Y. (2015). Is model fitting necessary for model-based fMRI? *PLoS Computational Biology*, *11*, 1–21.

Worthy, D. A., & Maddox, W. T. (2014). A comparison model of reinforcement-Learning and win-stay- lose-shift decision-making processes: A tribute to W.K. Estes. *Journal of Mathematical Psychology*, *59*, 41–49.

Yin, H. H., Ostlund, S. B., Knowlton, B. J., & Balleine, B. W. (2005). The role of the dorsomedial striatum in instrumental conditioning. *European Journal of Neuroscience*, *22*, 513–523.

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*, 681–

692.

Zeki, S., & Shipp, S. (1988). The functional logic of cortical connections. *Nature*, *335*, 311–317.

Acknowledgements

Firstly, I would like to thank my supervisor Prof. Dr. Marco Steinhauser for the best support any doctoral student could ever wish for. Without his dedication and commitment, I would not stand where I stand today and would probably never even have considered becoming a researcher in the first place. For the last five years, he was a constant source of inspiration and encouragement. His thoughtful comments and the stimulating discussions have substantially contributed to this thesis and steered me towards its completion. Despite his many obligations, he always made me feel appreciated and provided me with an environment to thrive as a young researcher. Marco, you are a true role model and a prudent mentor!

I am glad for all my colleagues. What would work have been without you? Thank you, Alodie, Clara, Daniela, Eva, Francesco Johanna, Julia, Kathrin, Klemens, Lukas, Martin, Miriam, Peter, Petra, and Steffi! I would like to give very special thanks to Dr. Robert Steinhauser and Dr. Benjamin Ernst for their indispensable feedback and discussion along the way.

I also would like to express my sincere appreciation to my collaborators from Ghent, who made a vital contribution to Study 1. Thank you, Wioleta Walentowska, Gilles Pourtois and Mario Carlo Severo.

Moreover, many thanks to the co-evaluator of my thesis, Prof. Dr. Michael Zehetleitner, who kindly agreed to assess my work.

Finally, I must not forget my friends and family. I am grateful to my mother, my father and my two sisters, for their unconditional love and their moral and emotional support throughout my entire life. I am also grateful to my friends who have accompanied me through thick and thin and always had my back. You all have made me who I am today.