Hierarchical Reinforcement Learning in Behavior and the Brain

José J. F. Ribas Fernandes

Dissertation presented to obtain the
Ph.D degree in Biology | Neuroscience

Instituto de Tecnologia Química e Biológica | Universidade Nova de Lisboa

Oeiras,
November, 2013

INSTITUTO
DE TECNOLOGIA
QUÍMICA E BIOLÓGICA
/UNL

Knowledge Creation

Hierarchical Reinforcement Learning in Behavior and the Brain

José J. F. Ribas Fernandes

Dissertation presented to obtain the
Ph.D degree in Biology | Neuroscience

Instituto de Tecnologia Química e Biológica | Universidade Nova de Lisboa

Research work coordinated by:

Oeiras,
November, 2013

To my mother

Acknowledgments

My deepest gratitude goes to my advisor, Matthew Botvinick. He is an exciting and inspiring person to work with. He is also a great mentor and has craftily nudged me in the direction of becoming a scientist. I would also like to thank Patrice Gensel, secretary of the Princeton Neuroscience Institute, and Leigh Nystrom, co-director of the Neuroscience of Cognitive Control Laboratory, for their successful efforts to circumvent Princeton's administrative hurdles for eternally visiting students.

I am immensely grateful to the dear friends I made in Princeton. Anna, Mike, Carlos, Judy, Matt, Natalia, and Wouter gave unconditional support when additional motivation was needed. They also made the time spent in bustling Princeton enjoyable, together with Alec, Amelia, Eric, Liliana, Jiaying, Mike (Todd), Richard, Sam, and Valeria.

My gateway to science was quite unique. I was fortunate to be part of the first year of the Gulbenkian/Champalimaud Neuroscience PhD Programme, where the faculty, Zach, Rui, Marta, Joe and Susana, conveyed the enjoyment of science in a gregarious and unassuming way. I am very grateful to Maria, Patrícia, Rodrigo, Margarida and Mariana, lifelong friends I made during the indelible year of classes in Lisbon and who were incredibly supportive notwithstanding the distance. I am also thankful to Tânia, Filipe, Bruno, Sara and Manuel, old friends with whom I have had the joy of sharing school and university with.

Last but never least, my family was crucial for the completion of my PhD. I deeply owe this achievement to my mother, who was unconditionally sweet and supportive, and would have been proud of my graduation, to my brother Manuel, to my father — from whom I inherited the neurotic skill of continuous criticism, to tia Fatinha, tio João, and tia Juca, to Teresa, and to my paternal grandparents and my grandmother.

Título

Aprendizagem por Reforço Hierárquico no Comportamento e no Cérebro

Resumo

A aprendizagem por reforço (*reinforcement learning*, RL) tem desempenhado um papel fundamental na compreensão da neurobiologia da aprendizagem e da tomada de decisão. Em particular, a associação entre a actividade fásica de neurónios dopaminérgicos no tegmento ventral e o erro na predição de recompensas (*reward prediction error*, RPE), quantificado segundo o algoritmo de diferenças temporais (*temporal-difference learning*, TD), constituiu uma descoberta chave na consolidação da relação entre neurociências e RL. Esta descoberta permitiu o avanço do conhecimento na distinção entre comportamento habitual e planeado, condicionamento, memória de trabalho, controlo cognitivo e monitoração de erros. Além destas contribuições, RL facilitou a compreensão dos défices cognitivos presentes na doença de Parkinson, depressão, défice de atenção e hiperactividade, e impulsividade.

No entanto, a maioria dos modelos de RL testados em neurociências tem uma capacidade limitada de aprendizagem de problemas complexos, nomeadamente à escala ecológica do comportamento humano. Esta restrição é um problema bem estudado em aprendizagem de máquinas, onde é conhecido como a maldição da dimensionalidade. Das várias soluções propostas, destacamos a aprendizagem por reforço hierárquico (*hierarchical reinforcement learning*, HRL) dada a prevalência da noção de hierarquia em psicologia e neurociências. Os métodos HRL facilitam a tomada de decisão e aprendizagem através da divisão hierárquica entre acções. Hierarquia neste contexto significa o parcelamento de acções subordinantes, que produzem recompensas (e.g., fazer café), em acções subordinadas (e.g., abrir a lata do café, aquecer água), uma caracteristica ubíqua do comportamento humano e animal.

A investigação apresentada nesta tese testou a hipótese que as estruturas responsáveis por RL estariam também envolvidas em HRL. Especificamente, que a actividade das áreas aferentes aos neurónios dopaminérgicos estaria associada a erros de predição ao nível de acções subordinantes (*pseudo-reward prediction errors*, PPEs).

Antes de investigar as respostas cerebrais a PPEs, uma série de estudos comportamentais, em humanos, procurou determinar se os resultados de acções subordinantes tinham uma influência nas escolhas de participantes diferente de recompensas primárias ou secundárias. Como previsto, os participantes escolheram com vista à maximização de recompensa, sem qualquer efeito de acções subordinadas. Este achado foi fundamental para excluir a possibilidade que erros na predição do resultado de acções subordinantes (PPEs) sejam RPEs. No entanto, de acordo com HRL, preferências por resultados de acções subordinantes foram reveladas quando os participantes se encontravam no momento de efectuar essa acção ou quando as escolhas não implicavam uma mudança de recompensa primária ou secundária.

Através de ressonância magnética funcional e electroencefalograma, em três estudos, foi demonstrado que actividade no córtex cingulado anterior (*dorsal anterior cingulate cortex*, dACC) esteve correlacionada com PPEs. Estas respostas reflectiram diferenças na magnitude, mas não no sinal, dos PPEs, em conformidade com o envolvimento desta área em aprendizagem por surpresa. Finalmente, um estudo adicional, com ressonância magnética funcional, procurou comparar directamente as respostas cerebrais a RPEs e PPEs. Foi encontrado que a actividade em dACC apenas reflectiu a magnitude, mas não o sinal, do erro de predição. No entanto, apenas se observaram respostas a RPEs e não a PPEs. Postulou-se que esta dissociação se tenha devido a competição no processamento de informação proveniente de acções que produzam recompensas finais e de acções subordinadas. Esta hipótese seria compatível com a primazia do efeito motivacional de acções

que produzam recompensa sobre acções subordinadas, em concordância com os estudos comportamentais referidos anteriormente. Em nenhum dos estudos de neuroimagem foram observadas respostas estriatais a PPEs ou a RPEs ao nível de accões subordinadas — apesar de ter sido replicado o efeito conhecido a RPEs monetários. Esta resposta selectiva de áreas aferentes de neurónios dopaminérgicos, e a dissociação observada no estriado entre RPEs em acções subordinadas e RPEs monetários, sugere que a dopamina não seja responsável por tomada de decisão em domínios hierárquicos.

Em conclusão, esta tese incita à inclusão de mecanismos hierárquicos nos modelos existentes de RL. Além desta extensão, permite o avanço do conhecimento da função de dACC, relacionando esta área com a tomada de decisão hierárquica.

Abstract

Reinforcement learning (RL) has provided key insights to the neurobiology of learning and decision making. The pivotal finding is that the phasic activity of dopaminergic cells in the ventral tegmental area during learning conforms to a reward prediction error (RPE), as specified in the temporal-difference learning algorithm (TD). This has provided insights to conditioning, the distinction between habitual and goal-directed behavior, working memory, cognitive control and error monitoring. It has also advanced the understanding of cognitive deficits in Parkinson's disease, depression, ADHD and of personality traits such as impulsivity.

However, the RL models that have mostly been tested in psychology and neuroscience do not scale well with the complexity of a learning and decision making problem, namely on the order of complexity present in ecological tasks. This is a well-studied problem in the machine learning literature, known as the curse of dimensionality. Out of the solutions that have been proposed to increase the scalability of RL mechanisms, one that is particularly appealing to psychology and neuroscience is hierarchical reinforcement learning (HRL). HRL exploits the task-subtask structure of sequential action, which is a ubiquitous feature of human and animal behavior.

The present research pursued the hypothesis that the same neural structures that are involved in RL are also also involved in HRL. In particular, that the activity of afferents of midbrain dopaminergic neurons should be sensitive to prediction errors at the level of subtasks, termed pseudo-reward prediction errors (PPEs). Before examining the neural correlates of PPEs, a set of behavioral studies confirmed that humans do not attach reward to subgoals, a crucial exploration to ensure that subgoal prediction errors are not RPEs. Nevertheless, in accordance with HRL, subgoal-related preferences were manifest when participants were engaged in a subtask and when their choices did not entail any change in reward. In three neuroimaging studies, using fMRI and EEG, activity in the dorsal anterior cingulate cortex

(dACC) correlated with PPEs. Moreover, dACC responded to differences in magnitude, but not valence, of the prediction errors. This is consistent with a role of dACC in learning through surprising events. A final fMRI study sought to compare the neural responses to PPEs to those of RPEs. Activity in dACC for prediction errors was again shown to be unsigned. However, responses were only observed for RPEs, and not PPEs. It is posited that this dissociation was the result of competition between information at the task and subtask level. This is compatible with the priority given to reward over any reinforcing effect of subtasks, which was observed in the behavioral studies. Across the reported studies, we observed no striatal engagement for PPEs, or for RPEs at the level of subtasks, though we replicated responses to monetary RPEs. The response of only a subset of dopaminergic afferents for PPEs and the striatal dissociation between subtask and monetary RPEs suggests that dopamine is not involved in hierarchical decision making.

In conclusion, this thesis encourages expansion of RL models in neuroscience to embrace mechanisms from HRL, and it advances the current understanding of dACC function, positing an involvement in hierarchical decision making.

Author Contributions

JJFRF and Matthew M. Botvinick designed the studies. JJFRF ran the studies, and analyzed the data, together with Joe T. McGuire (first fMRI experiment, chapter 3), Alec Solway (EEG experiment, chapter 2), and Carlos Diuk (model fitting for behavioral experiments, chapter 2). JJFRF and Matthew M. Botvinick wrote the manuscripts that are included in this thesis (Ribas-Fernandes, Niv, & Botvinick, 2011; Ribas-Fernandes, Solway, et al., 2011).

Financial Support

JJFRF was the recipient of a doctoral scholarship from FCT, reference SFRH/BD/33273/2007.

Contents

Chapter 1

Learning and Decision Making in Hierarchical Reinforcement Learning

1.1 Chapter Summary

This chapter lays out the fundamentals of Reinforcement Learning (RL) and expounds the need to move beyond the algorithms usually employed in the literature.[1]

- Temporal-difference RL has had a tremendous success in understanding the neural foundations of decision making and learning.
- Computer scientists have pointed out that model-free RL does not scale well with domain complexity. Psychologists and neuroscientists have urged for more scalable, and thus more plausible, models.
- Hierarchical Reinforcement Learning (HRL) ameliorates the scalability of model-free RL by introducing extended sequences of actions in the behavioral repertoire of an agent.
- The computational concept of hierarchy resonates with longstanding ideas in psychology and neuroscience.
- The neural correlates of model-free RL can be easily extended to yield putative neural mechanisms for HRL.

[1] Sections of this chapter were published in Ribas-Fernandes, Niv, and Botvinick (2011).

1.2 The Success of Model-free RL

Over the past two decades, ideas from computational reinforcement learning (RL) have had an important and growing effect on neuroscience and psychology. The impact of RL was initially felt in research on classical and instrumental conditioning (Barto & Sutton, 1981; Sutton & Barto, 1990; Wickens, Kotter, & Houk, 1995). Soon thereafter, its reach extended to research on midbrain dopaminergic function, where the temporal-difference (TD) learning paradigm provided a framework for interpreting temporal profiles of dopaminergic activity (Barto, 1995; Houk, Adams, & Barto, 1995; Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997; Niv, 2009, for a review).

Subsequently, actor-critic architectures for RL have inspired new interpretations of functional divisions of labor within the basal ganglia and cerebral cortex (for a review, see Joel, Niv, & Ruppin, 2002), and RL-based accounts have been advanced to address issues as diverse as motor control (e.g., Miyamoto, Morimoto, Doya, & Kawato, 2004), working memory (e.g., O'Reilly & Frank, 2006), performance monitoring (e.g., Holroyd & Coles, 2002), and the distinction between habitual and goal-directed behavior (e.g., Daw, Niv, & Dayan, 2005).[2] It has also advanced the understanding of cognitive deficits in Parkinson's disease, depression, ADHD and of personality traits such as impulsivity (e.g., Frank & Seeberger, 2004; Maia & Frank, 2011, for a review).

[2]Curiously, ideas from neuroscience have in turn inspired algorithmic approaches in the computational RL literature, namely the fact that phasic dopamine activity also reflects novelty (Singh, Barto, & Chentanez, 2005; Reed, Mitchell, & Nokes, 1996; Dayan & Balleine, 2002; Kakade & Dayan, 2002).

1.3 Fundamentals of RL

RL problems comprise four elements: a set of world states, a set of actions available to the agent in each state, a transition function, which specifies the probability of transitioning from one state to another when performing each action, and a reward function, which indicates the amount of reward (or cost) associated with each such transition. Given these elements, the objective of learning is to discover a policy, that is, a mapping from states to actions, that maximizes cumulative discounted long-term reward.

There are a variety of specific algorithmic approaches to solving RL problems (for reviews, see Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998; Szepesvari, 2010). We focus on the approach that has arguably had the most direct influence on neuroscientific translations of RL, referred to as the actor-critic paradigm (Barto, 1995; Joel et al., 2002). In actor-critic implementations of RL, the learning agent is divided into two parts, an actor and a critic, as illustrated in Figure 1.1A (for example, Barto, Sutton, & Anderson, 1983; Houk et al., 1995; Suri, Bargas, & Arbib, 2001; Joel et al., 2002). The actor selects actions according to a modifiable policy, $\pi(s)$ in Figure 1.1, which is based on a set of weighted associations from states to actions, often called action strengths. The critic maintains a value function, $V(s)$, which associates each state with an estimate of the cumulative, long-term reward that can be expected subsequent to visiting that state. Importantly, both the action strengths and the value function must be learned based on experience with the environment. At the outset of learning, the value function and the actor's action strengths are initialized, for instance, uniformly or randomly, and the agent is placed in some initial state. The actor then selects an action, following a rule that favors high-strength actions but also allows for exploration. Once the resulting state is reached and its associated reward is collected, the critic computes a TD prediction error, denoted δ in Figure 1.1. The value that was attached to

the previous state is treated as a prediction of the reward that would be received in the successor state, $R(s)$, plus the value attached to that successor state. A positive prediction error indicates that this prediction was too low, meaning that an outcome turned out better than expected. Of course, the reverse can also happen, yielding a negative prediction error.

The prediction error is used to update both the value attached to the previous state and the strength of the action that was selected in that state. A positive prediction error leads to an increase in the value of the previous state and the propensity to perform the chosen action at that state. A negative error leads to a reduction in these. After the appropriate adjustments, the agent selects a new action, a new state is reached, a new prediction error is computed, and so forth. As the agent explores the environment and this procedure is repeated, the critic's value function becomes progressively more accurate, and the actor's action strengths change so as to yield progressive improvements in behavior, in terms of the amount of reward obtained.

The actor-critic architecture, and the TD learning procedure it implements, have provided a very useful framework for decoding the neural substrates of learning and decision making. Although accounts relating the actor-critic architecture to neural structures do vary (for a review, see Joel et al., 2002), one influential approach has been to identify the actor with the dorsolateral striatum (DLS), and the critic with the ventral striatum (VS) and the mesolimbic dopaminergic system (see, for instance, O'Doherty et al., 2004; Daw, Niv, & Dayan, 2006, Figure 1.1B). Dopamine (DA), in particular, has been associated with the function of conveying reward prediction errors to both actor and critic (Barto, 1995; Montague et al., 1996; Schultz et al., 1997). This set of correspondences will provide an important backdrop for our later discussion of Hierarchical Reinforcement Learning (HRL) and its neural correlates.

1.4 The Curse of Dimensionality and the Blessing of Abstraction

The actor-critic framework, and other TD implementations, share the simplicity of not keeping a model of the environment. This comprises the transition and reward functions. Such algorithms are therefore model-free, in contrast to model-based RL, which explicitly learns and uses the transition probabilities for computing values.

The simplicity of model-free RL comes at a cost. As the number of states and actions increases, the time to reach an optimal policy increases exponentially (Bellman, 1957), as a large amount of visitations to each state-action pair is required to achieve a useful estimate of the value. This is a well-known problem in the computational literature and impacts the scalability of model-free RL. In spite of this problem, the testbeds of RL in neuroscience and psychology have mostly been tasks with small complexity, compared with the complexity of human behavior (Dayan & Niv, 2008; Daw & Frank, 2009). The poor scalability thus questions the validity of RL algorithms to human behavior.

Two computational concepts have been proposed to address the scaling problem, abstraction and generalization (a division according to Ponsen, Taylor, & Tuyls, 2010). In abstraction, the representation of the learning problem is changed to only include relevant properties to behavior. If the change is applied to states it is called *state* or *structural abstraction* (Li, Walsh, & Littman, 2006), and if it is employed in actions then it is termed *temporal abstraction* (Precup, 2000). As opposed to an abstracted representation, an unmodified representation is called *flat*. In contrast to abstraction, in generalization the representation of a learning problem is not changed. Instead, similarities between states or actions are leveraged.

These two ideas are combined into different sets of RL methods, 1. hierarchical reinforcement learning (HRL, Barto & Mahadevan, 2003; Hengst,

2012),using temporal, sometimes state, abstraction — whereby agents can use actions at different levels of abstraction, 2. transfer learning (Taylor & Stone, 2009), training on a source problem and applying knowledge to a target problem, using generalization, and 3. relational RL (Džeroski, De Raedt, & Driessens, 2001), which uses inductive logic to represent actions and states, and employs a mixture of abstraction and generalization.

The focus of this thesis is on HRL, where sequences of actions are represented according to a part-whole structure, illustrated in Figure 1.2. The choice of abstraction in the action domain is sustained by the ample work in psychology, pointing to a hierarchical structure of behavior and its neural representations (see the section on hierarchy in behavior, and Botvinick, 2008).

Temporal abstraction has been around since early work in artificial intelligence (Newell & Simon, 1972; Fikes, Hart, & Nilsson, 1972). In its inception, it involved using aggregated actions, called *macro-operators* to facilitate planning. Since then, work in AI focused on the representation of the macro-operators, learning the sequence of actions of the macro-operators, planning in stochastic environments, and finding useful subgoals, very much the same questions that are approached by HRL (for a review, see Precup, 2000).

The use of these temporally-extended, aggregated actions allows systems to solve problems in a smaller number of steps, as illustrated in Figure 1.3. Assuming these sequences are known and are appropriate to the goal in question, the problem of exploration becomes simpler. Many learning problems can be decomposed into smaller ones and it is often the case that the composing units are shared with other tasks (Newell & Simon, 1972). For instance, *drive* occurs both in *get groceries* and *get to airport* and it would be inefficient to learn the same sequence twice. In any case, it is important to ask about the origin and usefulness of these sequences (for simulations of useful and prejudicial sequences, see Jong, Hester, & Stone,

2008; Botvinick, Niv, & Barto, 2009, and the section on subgoal discovery in the last chapter).

1.5 Introduction to Hierarchical Reinforcement Learning

HRL has two main objectives, reward maximization while learning policies at several levels of abstraction, which this thesis focus on, and, to a lesser extent, determine which levels of abstraction are relevant for behavior. Formally speaking, the HRL setting is no longer a Markov decision process (MDP), but rather a semi-Markov decision process (SMDP), where dependencies are no longer between single transitions, but rather span sequences of states and actions, sometimes called histories.

Among HRL methods to learn policies hierarchically, the most popular are *options* (Sutton, Precup, & Singh, 1999), MAXQ (Dietterich, 1998) and HAM (hierarchy of abstract machines by Parr, 1998; Parr & Russell, 1998). As defined in Diuk and Littman (2008), MAXQ is an algorithm which receives a multi-level hierarchical task decomposition as an input, something that can be both powerful and limiting, and incorporates state abstraction at each level. HAM specifies a series of non-deterministic finite state machines, where "elements in HAMs can be thought of as small programs, which at certain points can decide to make calls to other lower-level programs" (Diuk & Littman, 2008). Both MAXQ and HAM can be expressed as options (Precup, 2000), and in general the options framework is the most parsimonious and the one that requires least extensions from flat, model-free RL (Sutton et al., 1999). For these reasons, Botvinick, Niv, and Barto (2009) proposed the options framework as the first approximation to understanding the neural correlates of reward-based hierarchical learning.

Other HRL methods differ from *options*, MAXQ or HAM in that they use state abstraction (Dayan & Hinton, 1993, though MAXQ also uses state

abstraction), target problems with partial observability (Wiering & Schmidhuber, 1998), address continuous-time MDPs (Ghavamzadeh & Mahadevan, 2001) or solve concurrent activities (Rohanimanesh & Mahadevan, 2001).

1.5.1 Options

In *options* the action space is extended to include temporally extended actions, called options,[3] illustrated in Figure 1.4, in addition to regular, primitive, actions. This parses the core MDP into smaller MDPs, each being a separate learning problem, with its own reward function (in terms of *pseudo-reward*). Regardless of the structure of the action space, the observable output of behavior is a sequence of primitive actions. In reality, primitive actions can be considered one-step options. However, for clarity, we will continue to refer to primitive actions as actions. In *options*, the state space in *options* is the same as the core MDP, at all levels of the hierarchy (differing from, for example, Dayan & Hinton, 1993).[4]

Options are characterized by three components: (1) a set of initiation states, determining at which states an option is available for selection, (2) a set of termination conditions, mapping states to a probability of termination, and (3) option-specific policies π_o. Option-specific policies can invoke primitive actions, or other options.

Top-level action selection and learning. Whether to select an option at a particular state is governed by values which reflect expected discounted sum of rewards, $V(s)$, similarly to selection of actions in regular RL, Figure 1.5. In options, however, $V(s)$ reflects the extended nature of

[3] We will use italic to denote the framework and regular type for the extended actions.

[4] The term "option" exists in other fields of psychology (Kalis, Kaiser, & Mojzisch, 2013), as a statement that is relevant to the attainment of a goal, not necessarily in the domain of actions (Ward, 2007), in the problem-solving literature as possible steps that can be taken for the attainment of an action (Klein, Wolf, Militello, & Zsambok, 1995), and in the context of motor decisions in sport, where an option is almost on the opposite end of temporal abstraction, describing sets of motor primitives such joint angles (Raab & Johnson, 2007).

the option:

$$V^{\pi}(s) = E\{r_{t+1} + ... + \gamma^{k-1} r_{t+k} + \gamma^{k} V^{\pi}(s_{t+k}) | o, s, t\} \tag{1.1}$$

Where k is the duration of option o, taken at state s_t, according to policy π, terminated in state s_{t+k} and discounted by γ. Notice that the value reflects option policies, as it is a sum of the yields during option o: $E\{r_{t+1} + ... + \gamma^{k-1} r_{t+k}\}$, with the value at the termination state $V^{\pi}(s_{t+k})$. This is illustrated in Figure 1.5.

Learning at the reward level is very similar to flat RL, and consists of regular and extended updates. Taking as an example the MDP shown in Figure 1.5, the first prediction error (the green arrow between s_1 and s_2) is equal between the hierarchical (top) and flat (bottom) agents, because in both cases a primitive action was selected:

$$V(s_1) \leftarrow V(s_1) + \delta, \delta = \alpha[r_2 + \gamma V(s_2) - V(s_1)] \tag{1.2}$$

Where α is the learning rate. In s_2, however, the update will be different. $V(s_2)$ is updated with extended reward prediction errors (the long green arrow in 1.5):

$$V(s_2) \leftarrow V(s_2) + \delta, \delta = \alpha[r_3 + \gamma r_4 + ... + \gamma^{k} V(s_5) - V(s_2)] \tag{1.3}$$

This update happens after the option has been terminated, s_5 and the agent has observed the entire sequence of accrued rewards. The next time an agent is in s_2, $V(s_2)$, which reflects an encapsulated prediction of rewards, will be used to select actions as flat values in regular RL (e.g., using softmax).

Option-level action selection and learning. Once an option is selected, option-specific values come into play (V_o). These only affect action selection *while* the agent is executing the option. Values V reflect expected discounted cumulative reward, whereas V_o reflects expected discounted cu-

mulative *pseudo-reward* (depicted with a yellow asterisk in Figure 1.6). This is a hedonic signal that is delivered at the last state of the option, called the subgoal, and used to drive learning of option policies, independently of the top level. This way the agent learns option policies, that can be later transferred across problems with similar task structure, as well as root-level policies. It is paramount that this is separate from reward, otherwise an agent would prematurely terminate its behavioral course at the subgoal.

Learning while an option is executed also resorts to TD learning. At the end of each action, the learning agent observes a certain amount of pseudo-reward and the option-specific value of the new state, and with this information a pseudo-reward prediction error is computed:

$$V_o(s_t) \leftarrow V_o(s_t) + \delta, \delta = \alpha[\psi r_{t+1} + \gamma V_o(s_{t+1}) - V_o(s_t)] \tag{1.4}$$

Where ψr designs the amount of pseudo-reward, and δ is the pseudo-reward prediction error, or PPE (depicted by the lower green arrows in Figure 1.6). Crucially these updates and quantities are independent of the top level, and do not exist in standard RL methods.

Options bears the most resemblance with the methods that have been tested in neuroscience (e.g., O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003). For this reason, we adopt the framework of Botvinick, Niv, and Barto (2009), who have proposed *options* as the parsimonious candidate for extending neural RL mechanisms to hierarchical domains.

1.6 Hierarchy in Action and its Neural Implementation

The aim of this section is to review evidence for a hierarchical organization behavior and its neural bases, and thus provide a scaffold for neural HRL. There are important arguments to keep in mind while discussing evidence of

hierarchy. There is no clear behavioral hallmark of hierarchy, unlike devaluation for goal-directed behavior (Dickinson, 1985). Only when cognitive paradigms became more refined was it possible to detect hierarchical structure in behavior (for example, Rosenbaum, Kenny, & Derr, 1983; Crump & Logan, 2010; Collins & Frank, 2013), looking at patterns of transfer, priming and switch costs. Secondly, any task with hierarchical structure can be solved by a flat agent, without abstract actions or abstract representations (Sutton et al., 1999; Botvinick & Plaut, 2004). Finally, there is a *utility problem* in adding temporally-extended sequences to control. This was recognized by early AI (Lehman, Laird, & Rosenbloom, 1996), and only recently has it received systematic attention (Jong et al., 2008; Van Dijk, Polani, & Nehaniv, 2011; Solway et al., submitted). In spite of hierarchy being historically difficult to detect, not strictly necessary, and sometimes prejudicial to learning, hierarchical behavior is ubiquitous, as we review below.

1.6.1 Hierarchical structure in behavior

Karl Lashley (1951) is credited with first pointing out the need for a non-sequential account of behavior. He argued that selecting an action at every behavioral transition was inefficient — thus coming closer to the computational justification for hierarchy, as presented in section 1.4. He supported this idea with the pattern of errors in language, which showed evidence of higher-level mental plans, instead of being the product of single stimulus-response associations. For example, when typing *groceries*, errors will often reflect a forthcoming letter or subsequence, *groc***r***eis*, rather than a random letter, *groc***k***eis*. From a computational perspective, this example suggests that errors depend on extended policies, very much like an option (though the *options* framework does not make a direct prediction about errors). Because of this earlier reliance on language, it took 40 years until a similar statement was published in the animal literature (Terrace, 1993; Fountain,

Wallace, & Rowan, 2002, though research on goal-directed action already mitigated a stimulus-response account of behavior).

After Lashley's argument, early work in cognitive psychology focused on how such hierarchical behavior could be generated, from a cognitive perspective and by mapping abstract modules of action directly to a control unit. A pioneering model in this regard was the TOTE model of Miller, Galanter, and Pribram (1960, Test-Operate-Test-Execute), where each unit resembled a finite-state machine as in HAM. As cited in Botvinick (2008), this was followed by research on scheduling of control units (in memory, Estes, 1972; typing and speech, Rumelhart & Norman, 1982; Mackay, 1987; and in the domain of everyday action, Cooper & Shallice, 2000), on the combination of habitual and supervisory units (Norman & Shallice, 1986), on models with biologically inspired units (Dehaene & Changeux, 1997; Grossberg, 1986; Houghton, 1990), and on more abstract proposals of hierarchical structure in action performance and understanding (Schank & Abelson, 1977). Beyond providing a generative model for hierarchical behavior, these models also gave support to Lashley's suggestion that errors are a result of higher-level plans (e.g., Cooper & Shallice, 2000).

An important change of paradigm was introduced by connectionist models (Elman, 1990; Cleeremans, 1993; and later followed by Botvinick & Plaut, 2004; Botvinick, 2007; Frank & Badre, 2012). Contrary to prior models, hierarchical representations were not explicitly built in the model. Rather, these were represented in the patterns of weights between hidden units, and arose through learning, without input from the user. The fact that behavior can be achieved through very different representations highlights the important possibility that the task structure might not be mirrored in the actual neural implementation (Uithol, van Rooij, Bekkering, & Haselager, 2012).

Early on, evidence for hierarchical structure in behavior were thorough registrations of slips of action, in line with Lashley's language examples,

drawing from research on verbal behavior (Garnham, Shillcock, Brown, Mill, & Cutler, 1981), and routine actions, normal and pathological (Reason, 1979; Schwartz, Reed, Montgomery, Palmer, & Mayer, 1991; Humphreys & Forde, 1998). Later, springing from a renewed focus on hierarchical behavior, other sources of research provided evidence to hierarchy in behavior: research on event perception (Zacks & Tversky, 2001; Kurby & Zacks, 2008) — showing how people can parse streams of actions into meaningful subsequences; typing (Logan, 2011) — showing priming and Stroop-like effects at different levels of abstraction; and developmental psychology (Saffran & Wilson, 2003; Whiten, Flynn, Brown, & Lee, 2006) — showing how infants learn simultaneously at different levels of abstraction. Direct behavioral evidence for human learning at several timescales according to RL principles comes from recent neuroimaging studies (Haruno & Kawato, 2006; Diuk, Tsai, Wallis, Botvinick, & Niv, 2013).

In the animal literature, evidence came from chunking of action sequences and analysis of grooming sequences in rodents (Fentress, 1972; Berridge, Fentress, & Parr, 1987, in a similar vein to the earlier descriptive analyses, e.g., Reason, 1979) — which demonstrates the existence of temporally extended policies; and list learning (Terrace, 1993), in pigeons and monkeys — eliciting similar errors to Lashley's misinsertions (for a review, see Conway & Christiansen, 2001).

Even though state abstraction is not part of many HRL methods, we should mention studies that involve this type of abstraction. This is based on the fact that the two abstractions might share many of the prefrontal substrates, as discussed in the next section, and that, from an ecological perspective, state and temporal abstraction often co-occur. In this setting, research on task sets comes to bearing (MacLeod, 1991; Monsell, 2003), showing that people learn abstract rules, and that errors and priming effects are dependent on which abstract rule is control of behavior — though research in this field has mostly concentrated on the dynamics of task switch-

ing. Particularly relevant are studies that show abstraction even in settings where it is not necessary to do so (Badre, Kayser, & D Esposito, 2010; Frank & Badre, 2012; Collins & Frank, 2013, the later two studies combining neuroimaging and behavior). SimiIarly to effects of task sets in humans, context effects in conditioning in rodents show that single actions depend on more abstract states (Courville, Daw, & Touretzky, 2006; Gershman & Niv, 2012).

1.6.2 Neural implementation of hierarchical sequential action

The structures that have figured in action selection and performance in hierarchical domains have been the dorsolateral prefrontal and orbitofrontal cortices (DLPFC and OFC), dorsolateral striatum (DLS), and to a lesser extent, the ventral striatum (VS).[5] For the reason that theoretical understanding and empirical evidence are still growing, rather than adopting a specific framework to describe the activity of these areas, we review ways in which neural responses differ from a flat representation. Only in the next section do we discuss how these areas can be associated with a neural instantiation of *options*.

The more straightforward and earliest forms of implementation of task hierarchies assumed that the action hierarchy was mirrored in the control hierarchy, where each unit reflected a subtask which would be sequentially activated, as in Figure 1.7A (e.g., Miller et al., 1960; Cooper & Shallice, 2000). However, even though neuroanatomical hierarchical divisions might be obvious (e.g., Goldman-Rakic, 1987), representations might not contain any direct elements of the action hierarchy (Botvinick & Plaut, 2004; Botvinick, 2007; Reynolds & Mozer, 2009; Uithol et al., 2012), as proved by connectionist models. The exact same behavior can be produced with-

[5] One source of research that we do not mention is literature on perception of goals, which involves the inferior parietal sulcus (Hamilton & Grafton, 2006; Bonini et al., 2011).

out any explicit division of labor (Figure 1.7B, Botvinick & Plaut, 2004; or Koechlin, Ody, & Kouneiher, 2003, *vs.* Reynolds & Mozer, 2009), and even if there is a hierarchical separation between units, each might not map onto particular subactions (Botvinick, 2007). To add to the confusion, hierarchical divisions of labor might be beneficial even for non-hierarchical tasks (Botvinick, 2007).

Prefrontal cortex. Dorsolateral prefrontal cortex (DLPFC, BA 9, and 46) has been extensively studied in humans and nonhuman primates, in lesion and normal studies. A single pattern of DLPFC activation has been associated with an entire mapping from stimuli to responses (Hoshi, Shima, & Tanji, 1998; White & Wise, 1999; Asaad, Rainer, & Miller, 2000; Shimamura, 2000; Wallis, Anderson, & Miller, 2001; Bunge, 2004; Rougier, Noell, Braver, Cohen, & O'Reilly, 2005; Johnston & Everling, 2006), and not the details of the task itself, corroborating the guided activation theory (Miller & Cohen, 2001). DLPFC has also been found to code for progression in a multistep task (Hasegawa, Blitz, & Goldberg, 2004; Knutson, Wood, & Grafman, 2004; Amiez & Petrides, 2007; Berdyyeva & Olson, 2010; Saga, Iba, Tanji, & Hoshi, 2011) and action sequence boundaries (Fujii & Graybiel, 2003; Farooqui, Mitchell, Thompson, & Duncan, 2012), a type of response that has also been found in dorsolateral striatum.

The function of DLPFC is often considered together with that of frontopolar cortex (BA 10) and anterior premotor cortex (BA 8) in a number of theories which posit a rostrocaudal allocation of function. Each theory focus on a particular variable: amount of information required to reduce response uncertainty (Koechlin & Summerfield, 2007), level of state abstraction (Badre & D'Esposito, 2007; Badre, Hoffman, Cooney, & D'Esposito, 2009), temporal abstraction (Sirigu et al., 1995; Fuster, 1997; Grafman, 2002; Wood & Grafman, 2003, 2003; Zalla, Pradat-Diehl, & Sirigu, 2003), relational complexity (Christoff, 2003; Christoff & Keramatian, 2007), or domain specificity (Sakai & Passingham, 2006; Courtney, Roth, & Sala,

2007) — for reviews, see Hoshi (2006), Botvinick (2008), Badre (2008), and Badre and D'Esposito (2009). The exact contribution of each area and the nature of the gradient is still a subject of controversy (Reynolds, O'Reilly, Cohen, & Braver, 2012; Duncan, 2013).

One possible source of confusion for theories of lateral prefrontal cortex, is that abstract actions are often associated with multiple effectors, as well as abstract, multimodal states. In addition, different cognitive processes might be recruited for each level of abstraction (e.g., temporally abstract actions require working memory, whereas primitive actions do not), making it that the organizing principle might not be about levels of hierarchy, but cognitive processes.

Neurophysiological data has shown that within OFC (BA 11, 13, and 14) reward-predictive activity tends to be sustained, spanning temporally extended segments of task structure (Schultz, Tremblay, & Hollerman, 2000). In addition, the response of OFC neurons to the receipt of primary rewards has been shown to vary depending on the wait-time leading up to the reward (Roesch, Taylor, & Schoenbaum, 2006).

Another prefrontal area that has been involved in hierarchical behavior is the pre-supplementary motor area (pre-SMA, BA 8). In addition to the putative role at the lower levels of hierarchy as stipulated by rostro-caudal gradient theories, this area has been found to code for sequences of movement as a whole (Shima, Mushiake, Saito, & Tanji, 1996; Nakamura, Sakai, & Hikosaka, 1998; Shima & Tanji, 2000; Bor, Duncan, Wiseman, & Owen, 2003; Kennerley, Sakai, & Rushworth, 2004; Averbeck & Lee, 2007; Shima, Isoda, Mushiake, & Tanji, 2007), and task set identity (Rushworth, Walton, Kennerley, & Bannerman, 2004).

Striatum. The dorsolateral striatum (DLS) has been shown to respond to the serial order of action in a sequence, but not of the action in isolation (in rodents, Aldridge, Berridge, Herman, & Zimmer, 1993; Aldridge & Berridge, 1998; Cromwell & Berridge, 1996; non-human primates, Kermadi,

Jurquet, Arzi, & Joseph, 1993; Kermadi & Joseph, 1995; Mushiake & Strick, 1995; Ravel, Sardo, Legallet, & Apicella, 2006). In addition, DLS has been shown to respond to the start and beginning of a sequence, something known as task bracketing (evidence coming mostly from rodents Jin & Costa, 2010; Barnes et al., 2011). This phenomenon might have a role in sequence chunking (Graybiel, 1998; Burkhardt, Jin, & Costa, 2009), such that lesions of DLS lead to impairments in building extended behavioral repertoires (Boyd et al., 2009; Tremblay et al., 2010). In addition, it is noteworthy that DLPFC projects heavily onto DLS (Alexander, DeLong, & Strick, 1986; Parent & Hazrati, 1995), thus consolidating the idea that these two structures are involved in sequential learning and selection — directly compared in Fujii and Graybiel (2005). These connections have supported the detailed computational models of Frank and Claus, which show how frontal inputs to the striatum could switch among different stimulus-response pathways (Rougier et al., 2005; Frank & Claus, 2006; O'Reilly & Frank, 2006).

Ventral striatum (VS) has figured less in hierarchical representation of behavior. Instead, it has been proposed to be involved in learning at different levels of abstraction (Ito & Doya, 2011). Consistent with a role in learning at multiple levels, a recent study has shown that ventral striatal codes for prediction errors at multiple levels of abstraction (Diuk et al., 2013) — at one level associated with deviations of outcomes from a bandit task, and at a higher level, deviations from outcomes of a sequence of bandits.

Finally, outside PFC and striatum, Daw, Courville, and Touretzky (2003) have suggested that DA responses are driven by representations which divide event sequences into temporally-extended segments, based on the pattern of responses to delayed rewards.

1.7 Extending Neural RL Mechanisms to HRL

HRL requires several computational extensions to regular RL: (1) Extended PEs, (2) Extended values, (3) Option-specific policies, (4) Option-specific value functions and pseudo-reward, and (5) Option-specific PEs or pseudo-reward prediction errors (PPEs). Botvinick, Niv, and Barto (2009) have proposed a mapping between these extensions and particular neural structures.

Other relevant, though less general, neural implementations of HRL have been put forth by Ito & Doya (2011) and Frank & Badre (2012) — see the last chapter for discussion of the differences between approaches. These have focused on cortico-striatal loops. Ito & Doya has proposed that higher levels of abstraction are represented more rostrally and medially in the basal ganglia, and more rostrally in the prefrontal cortex. Frank & Badre simulated and tested an extension of the working memory model of prefrontal-basal interactions for state abstraction (O'Reilly & Frank, 2006). Another implementation of HRL, Holroyd and Yeung (2012), has given a key role in option selection and maintenance to the medial frontal cortex, and to the interaction with dorsolateral and orbital frontal cortices.

1.7.1 Extended PEs

One important change in how PEs are computed is that HRL widens the scope of the events that the prediction error addresses. In standard RL, the prediction error indicates whether outcomes went better or worse than expected since the immediately preceding single-step action. In contrast, the prediction errors associated with *options* are framed around temporally extended events.

The widened scope of the prediction error computation in HRL resonates with work on midbrain DA function. In articulating this account, Daw et al. (2003) provided a formal analysis of DA function that draws on precisely

the same principles of temporal abstraction that also provide the foundation for HRL, namely an SMDP framework. Consistent with the involvement of dopamine in computing extended PEs, Diuk et al. showed ventral striatal responses to extended PEs, in addition to regular PEs.

1.7.2 Extended values

Note that in HRL, in order to compute a prediction error when an option terminates, certain information is needed. In particular, the critic needs access to the reward prediction it made when the option was initially selected, and for purposes of temporal discounting it also needs to know how much time has passed since that prediction was made. These requirements of HRL resonate with data concerning the OFC (Schultz et al., 2000; Roesch et al., 2006), which have shown that reward-predictive activity is sensitive to task structure.

1.7.3 Option-specific policies

As mentioned above, options come with their own policies, π_o in Figure 1.1C, assembled from a behavioral repertoire of actions and other options. This consists of two key variables: a representation of the identity of the option currently in control of behavior, and the sequence that is about to be performed, an option-level policy.

From a neuroscientific point of view, the representation of option identities seems very closely related to that commonly ascribed to the DLPFC. Prefrontal representations are not thought to implement policies directly, but instead select among stimulus-response pathways implemented outside the prefrontal cortex (Miller & Cohen, 2001). This division of labor fits well with the distinction in HRL between an option's identifier and the policy with which it is associated, which might be mapped onto DLPFC and preSMA/DLS respectively.

Research on frontal cortex also accords well with the stipulation in HRL that temporally abstract actions may organize into hierarchies, with the policy for one option (say, an option for making coffee) calling other, lower-level *options* (say, *options* for adding sugar or cream). This fits with the accounts suggesting that the frontal cortex serves to represent action at multiple, nested levels of temporal structure (Sirigu et al., 1995; Fuster, 1997; Grafman, 2002; Wood & Grafman, 2003, 2003; Zalla et al., 2003), possibly in such a way that higher levels of structure are represented more anteriorly (Botvinick, 2008; Badre, 2008; Badre & D'Esposito, 2009).

As reviewed earlier, neuroscientific interpretations of the basic actor-critic architecture generally place policy representations within the DLS. It is thus relevant that such regions as the DLPFC, SMA, pre-SMA and PMC — areas potentially representing *options* — all project heavily to the DLS (Alexander et al., 1986; Parent & Hazrati, 1995).

In HRL, as in guided activation theory, temporally abstract action representations in frontal cortex select among alternative (i.e., option-specific) policies. In order to support option-specific policies, the DLS would need to integrate information about the currently controlling option with information about the current environmental state, as is indicated by the arrows converging on the policy module in Figure 1.1.

Unlike the selection of primitive actions, the selection of *options* in HRL involves initiation, maintenance and termination phases. At the neural level, the maintenance phase would be naturally supported within DLPFC, which has been extensively implicated in working memory function (Postle, 2006; Courtney et al., 2007; D'Esposito, 2007). With regard to initiation and termination, it is intriguing that phasic activity has been observed, both within the DLS and in several areas of frontal cortex, at the boundaries of temporally extended action sequences (Zacks et al., 2001; Fujii & Graybiel, 2003; Morris, Arkadir, Nevet, Vaadia, & Bergman, 2004). Since these boundaries correspond to points where new *options* would be selected, boundary-

aligned activity in the DLS and frontal cortex is also consistent with a proposed role of the DLS in gating information into prefrontal working memory circuits (Rougier et al., 2005; O'Reilly & Frank, 2006).

1.7.4 Option-specific value functions and pseudo-reward

Another difference between HRL and ordinary TD learning is that learning in HRL occur at all levels of task structure. This is because, as mentioned in the section on HRL, there is a separate reward signal, noted pseudo-reward. The possible neural correlates for pseudo-reward are the structures that are posited to carry reward signals (Wise, 2002), the hypothalamus, and the pedunculopontine nucleus. The hypothetical neural correlate of such encapsulated value function would be the OFC (as reviewed in the section on extended value functions).

1.7.5 Option-specific PEs or pseudo-reward prediction errors (PPE)

At the topmost or root level, prediction errors signal unanticipated changes in the prospects for primary reward. However, in addition, once the HRL agent enters a subroutine, separate prediction error signals indicate the degree to which each action has carried the agent toward the currently relevant subgoal and its associated pseudo-reward. Note that these subroutine-specific prediction errors are unique to HRL.

In what follows, we refer to them as pseudo-reward prediction errors (PPE), reserving reward prediction error (RPE) for prediction errors relating to reward. Because the PPE is not found in ordinary RL, it can be considered a functional signature of HRL. If the neural mechanisms underlying hierarchical behavior are related to those found in HRL, it should be possible to uncover a neural correlate of the PPE. On grounds of parsimony, one would expect to find PPE signals in the same structures that

have been shown to carry RPE-related signals, in particular targets of midbrain dopaminergic projections including VS (Pagnoni, Zink, Montague, & Berns, 2002; O'Doherty et al., 2004; Hare, O'Doherty, Camerer, Schultz, & Rangel, 2008), anterior cingulate cortex (Holroyd & Coles, 2002; Holroyd, Nieuwenhuis, Yeung, & Cohen, 2003), as well as the habenula (Ullsperger & von Cramon, 2003; Matsumoto & Hikosaka, 2007; Salas & Montague, 2010) and amygdala (Breiter, Aharon, Kahneman, Dale, & Shizgal, 2001; Yacubian et al., 2006).

1.8 Aims: Exploring Neural Correlates of PPEs

The present work has focused on evidence for subtask-bounded prediction errors or pseudo-reward prediction errors (PPEs). The aims of the thesis are to:

- Develop a hierarchical paradigm where PPEs can be safely dissociated from RPEs (chapter Decision making in subtasks).
- Assess the influence of pseudo-reward on behavior (chapter Decision making in subtasks).
- Explore the neural correlates of positive and negative PPEs separately (chapter Neural correlates of pseudo-reward prediction errors).
- In a single paradigm, compare neural responses to PPEs and RPEs (chapter Neural correlates of pseudo-reward and reward prediction errors).

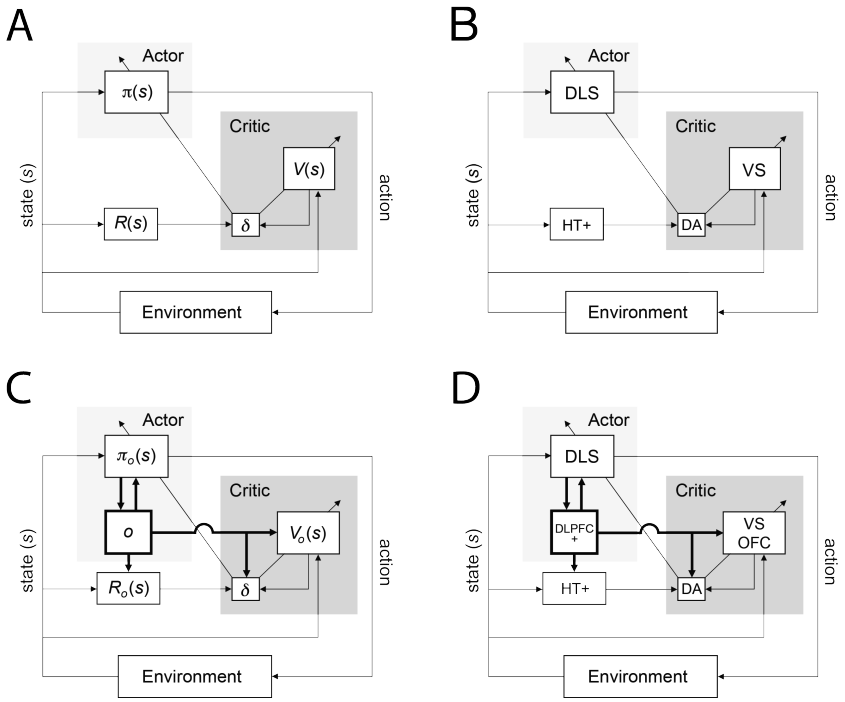

Figure 1.1. Fundamentals of the actor-critic architecture. (A) Relationship between agent and environment ($\pi(s)$ - policy, $V(s)$ - value at state s, $R(s)$ - reward at state s, δ - reward prediction error). Arrows represent direction of computations. (B) Neural correlates of actor-critic (DLS - dorsolateral striatum, DA - midbrain dopamine, VS ventral striatum, HT+ - hypothalamus and related reward structures, e.g., peduncunlopontine nucleus). (C) Extensions of actor-critic for options (o - option identifier, $\pi_o(s)$ - policy, $V_o(s)$ - option-specific value function, $R(s)$ - reward function, δ - pseudo-reward prediction error). (D) Putative neural extensions of the actor-critic architecture for options (DLPFC - dorsolateral prefrontal cortex, OFC - orbitofrontal cortex). From Botvinick, Niv, and Barto (2009).

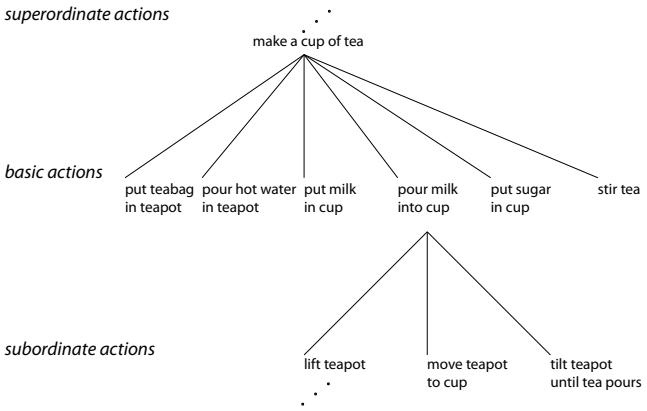

Figure 1.2. Hierarchical decomposition of the task of making tea. From Botvinick (2007), adapted from Humphreys and Forde (1998).

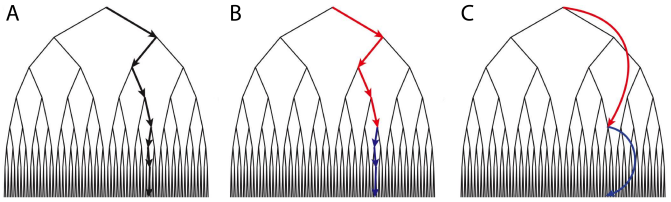

Figure 1.3. Temporal abstraction ameliorating the scalability of RL. In this Markov decision problem (MDP) an agent has to perform six sequential binary decisions. Only one of the branches yields reward. The flat agent (A), which uses only primitive actions, has to make six decisions. Assuming the agent has previously learned the red and blue sequences of actions (B), then exploration is greatly facilitated (C). This beneficial effect assumes that these sequences are already learned and appropriate for the domain in question — see the section on subgoal discovery for a discussion of this issue. From Botvinick, Niv, and Barto (2009).

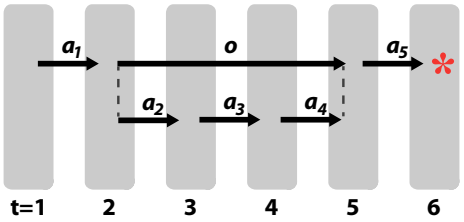

Figure 1.4. Behavioral repertoire of an *options* agent. Each grey box represents a state. The state at $t = 6$ yields reward, marked by the red asterisk. a denotes primitive actions and o, an option. The final sequence of behavior is the sequence of primitive actions *a1-5*. Adapted from Botvinick, Niv, and Barto (2009).

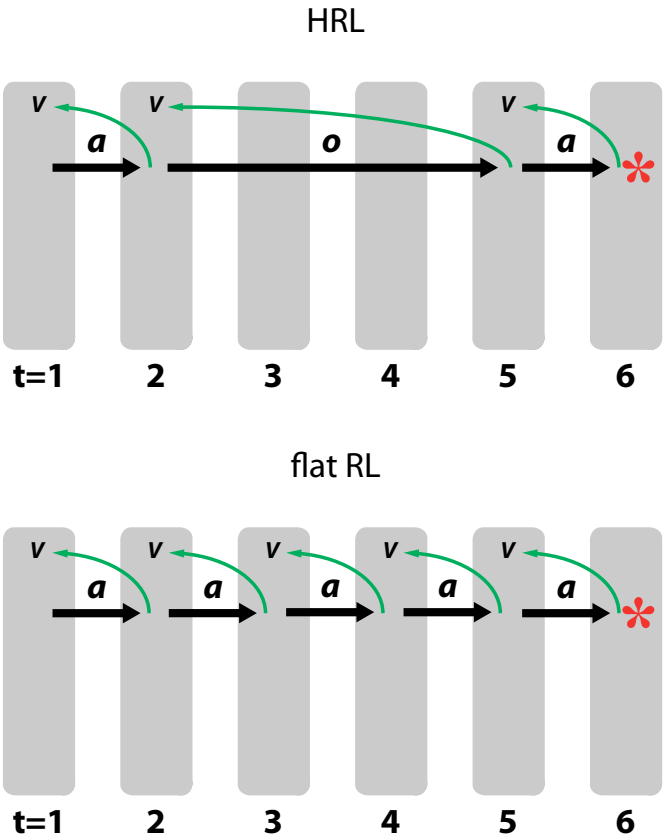

Figure 1.5. Reward-driven updates in HRL (top) and RL (bottom). Green arrows represent prediction errors. Reward-driven prediction errors in HRL can be exactly the same as in flat RL, as between s_1 and s_2, or reflect the extended nature of options illustrated in the long arrow between s_2 and s_5. Both values reflect an expected sum of discounted reward (marked by the asterisk), though with different temporal structure — see main text. Adapted from Botvinick, Niv, and Barto (2009).

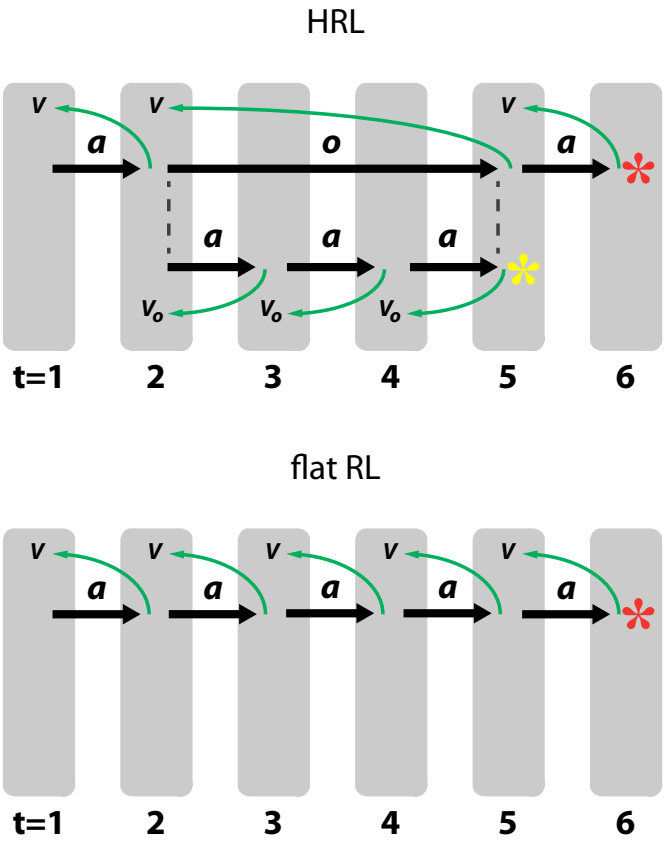

Figure 1.6. All learning updates in HRL and flat RL. In addition to learning values driven by reward, an HRL agent learns simultaneously at the subtask level. This is driven by pseudo-reward, and involves reward-independent updates called pseudo-reward prediction errors (PPE, represented by the lower-facing green arrows). PPEs are then used to update option-specific values V_o. Adapted from Botvinick, Niv, and Barto (2009).

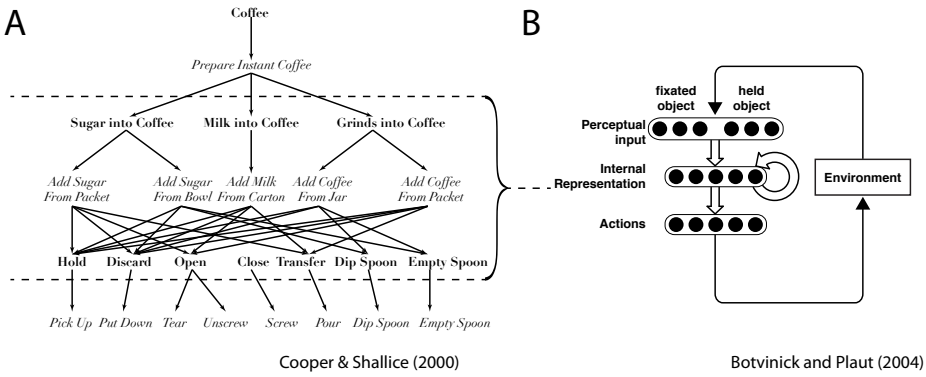

Figure 1.7. Comparison of symbolic and connectionist models of routine sequential action. (A) In the symbolic model of Cooper and Shallice (2000), the generative model of routine actions was composed of units which mirrored the parcellation of the task. (B) In the connectionist approach, Botvinick and Plaut (2004) modeled the same task with a recurrent neural network, without an explicit division of labor between units in the internal representation. The comparable parts of the two models are highlighted by the dashed lines.

Chapter 2

Decision Making in Subtasks

2.1 Chapter Summary

In this chapter we describe a series of experiments that examine behavioral predictions of HRL.

- A first experiment aimed at testing the relative influence of goals and subgoals on choice behavior. We designed a hierarchical spatial navigation paradigm where participants had to navigate a truck to pick up an envelope and then deliver it to a house. In this task, there was a clear incentive at minimizing the distance traveled. We offered two envelopes, trading-off action costs to attain the subgoal with those of the goal. Participants showed clear avoidance of goal costs and were indifferent to subgoal costs.

- A second and third experiment were variants of the hierarchical task, used to test predictions of HRL. The predictions are that subgoal preferences should be manifest when a participant is executing an option and that, the effect of such preferences should be larger when a choice is offered between subgoals of equal costs of goal attainment. There was a strong influence of goal preferences, as in the first experiment. Surprisingly, the choice patterns showed no influence of subgoals, suggesting that participants might have terminated the option "get subgoal" at the moment of choice (assuming a hierarchical representation).

- On a fourth experiment, we refined the previous paradigms using a minimal amount of pause, and voluntary, instead of forced, choice, intending to cause the least amount of disruption to option maintenance. We observed a clear influence of distance to subgoal in participants' preferences. This was distributed on a spectrum of preferences: approximately one third of the participants minimized the costs in the

first subtask, another third minimized the subgoal costs for a second subtask, and the remaining third was indifferent.

- The behavioral findings obtained appear consistent with primary predictions from HRL, in spite of an unexpected pattern of choices.

2.2 A Task Paradigm for Studying Hierarchical Decision Making

The behavioral experiments described in this chapter aim at tapping into HRL-like decision-making mechanisms.[1] This means that preferences should reflect the influence of reward, as is the case with flat agent, and, under certain conditions, reflect the influence of pseudo-reward, at the level of subgoals. When there is a trade off between reward and pseudo-reward, the former should completely dominate the latter. This is because there should be no attachment of reward to a subgoal (which would be a variant of a flat agent). This first experiment examines the prediction of goal dominance. Nevertheless, as stipulated by HRL, subgoal preferences should be revealed when there is no trade off between reward and pseudo-reward, or the subject is executing an option — something we explore in the ensuing experiments. There have been early behavioral studies in rodents examining the very same questions we pose here (Gilhousen, 1940; Spence & Grice, 1942; Kendler, 1943). In these studies, it is found that rats prefer closer subgoals independently of overall distance. However, upon closer inspection, the manipulation of subgoal distance also implied a change in goal distance. To our knowledge, there is no work addressing these questions of hierarchical preferences.

We designed a hierarchical paradigm based on a benchmark task from the computational HRL literature (the taxi task, Dietterich, 1998), the courier task. Participants played a video game which is illustrated in Figure 2.1. As detailed below, in spatial paradigms the distances to the goal and subgoal can be independently manipulated, which will prove crucial to determine whether people attach reward to the subgoal. Only the colored elements in the figure appeared in the task display. The overall objective of

[1] This experiment has been published in Ribas-Fernandes, Solway, et al. (2011), and some of the text and figures are adapted from this source.

the game was to complete a delivery as quickly as possible, using joystick movements to guide the truck first to the package and from there to the house. It is self-evident how this task might be represented hierarchically, with delivery serving as the externally rewarded, top-level goal and acquisition of the package as an obvious subgoal. This observation is not meant to suggest that the task must be represented hierarchically. Indeed, it is an established point in the HRL literature that any hierarchical policy has an equivalent non-hierarchical or flat policy, as long as the underlying decision problem satisfies the Markov property. For an HRL agent, delivery would be associated with primary reward and acquisition of the package with pseudo-reward. However, as mentioned in the introduction, pseudo-reward does not trade off with reward at the top level. For an RL agent, only delivery would be associated with reward, unless an agent attached reward to both pick-up and delivery of the package, which would show independent approach behavior.

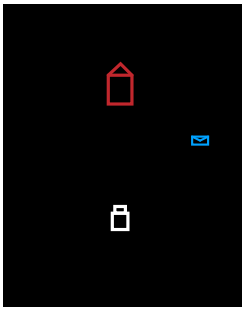

Figure 2.1. Hierarchical spatial paradigm. Participants had to pick up the package and deliver it to the house, using a joystick. Elements in the figure are not to scale.

Let us examine how action costs to the subgoal can be dissociated from those to the goal. Consider the envelope shown in Figure 2.2A. Any point on the solid line will have the same distance d_1 to the truck as the reference envelope. In Figure 2.2B, any point on the ellipse, the dashed line, will

have the same distance to the house, $d_1 + d_2$, shown by point **P**, but a different distance to the subgoal. Assuming action costs are proportional to the distance, we can then offer choices with independent subgoal and goal costs, and observe participants' preferences.[2]

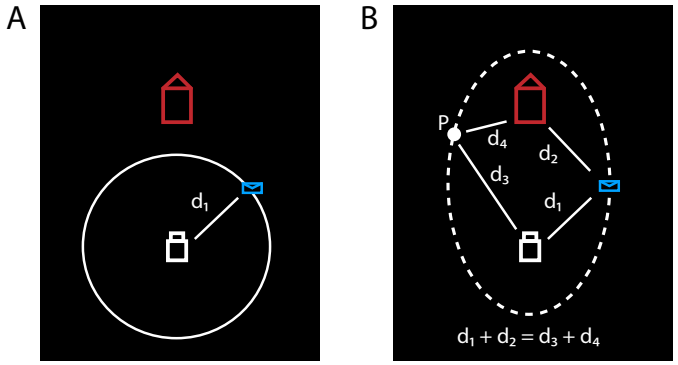

Figure 2.2. Dissociating action costs to attain the subgoal and the goal. (A) Costs to the subgoal. Any point of the solid circle will have the same action costs as the shown subgoal. (B) Costs to the goal. The costs to the goal are $d_1 + d_2$. By definition, any point on the dashed ellipse has the same costs to attain the goal as the envelope on the right (e.g., point **P**).

Methods

Participants. A total of 22 participants were recruited from the Princeton University community ($M = 20.3$, $SD = .5$, 11 male). All provided informed consent and received a nominal payment.

Task and procedure. Participants sat at a comfortable distance from a computer display in a closed room. A joystick was held in the right hand

[2]The "birds-eye" view of the display affords information about future states, which is different from the first-person perspective of a model-free agent in a gridworld. However, we wanted to make sure that both goal and subgoal distances were available to the participants at all times. Otherwise any manipulation of subgoal distance would be confounded with incomplete information about overall distance. In addition, as described in the next chapters, the manipulations eliciting prediction errors had elements of unpredictability which are independent of the agent's model of the task.

(Logitech International, Romanel-sur-Morges, Switzerland). The computerized task was coded using MATLAB (The MathWorks) and the MATLAB Psychophysics Toolbox, version 3 (Brainard, 1997). On each trial, three display elements appeared: a truck, an envelope and a house (Figure 2.1). Each joystick movement displaced the truck a fixed distance of 50 pixels. The orientation of the truck was randomly chosen after every such translation, and participants were required to tailor their joystick responses to the truck's orientation, as if they were facing its steering wheel (Figure 2.3). For example if the front of the truck were oriented toward the bottom of the screen, a rightward movement of the joystick would move the truck to the left. This aspect of the task was intended to ensure that intensive spatial processing occurred at each step of the task, rather than only at the beginning of a trial. Responses were registered when the joystick was tilted beyond half its maximum displacement (Figure 2.3A). Between responses the participant was required to restore the joystick to a central position (Figures 2.3A).

The experiment was composed of three phases. In the first phase, participants completed ten deliveries. At the beginning of each trial, the locations of the truck, envelope and house were determined randomly, with the constraint of being at least 100 pixels (two optimal steps) from each other (on a screen with resolution 1024 x 768 pixels). When the truck passed within 30 pixels of the envelope, the envelope would appear inside the truck and be carried within up to the delivery in the house. After picking up the envelope, when the truck passed within 35 pixels of the house, the truck would be shown inside the house and the message "Congratulations!" appeared for 300 ms.

The second phase consisted of ten further delivery trials. However, here, at the onset of each trial, the participant was required to choose between two packages (Figure 2.4). The location of the truck and the house was chosen randomly. The location of one package, designated subgoal one, was

randomly positioned along an ellipse with the truck and house as its foci and a major/minor axis ratio of 2. The position of the other package, subgoal two, was randomly chosen, subject to the constraint that it fell at least 100 pixels from each of the other icons. About one third of this second package, fell inside the ellipse.

At the onset of each trial, each package would be highlighted with a change of color, twice (in alternation with the other package, and counterbalanced across trials), for a period of 6 s (1.5 s for each package, twice). During this period the participant was required to press a key to indicate his or her preferred package when that package was highlighted. After the key press, the chosen subgoal would change to a new color. At the end of the choice period, the unchosen subgoal was removed, and participants were expected to initiate the delivery task. Importantly, participants had to wait 6s regardless of how fast they chose. The remainder of each trial proceeded as in phase one.

The third and main phase of the experiment included 100 trials. One third of these, interleaved in random order with the rest, followed the profile of phase two trials. The remaining trials began as in phase two but terminated immediately following the package-choice period. It should be noted that the termination at choice would not make an RL agent value the subgoal, as termination and choice were independent and choice happened before subgoal attainment. Participants were told that the first two parts of the experiment were intended to become acquainted with playing and choosing. In addition, they were also told that their choices had no influence on whether a trial would continue beyond choice.

Data analysis. To determine the influence of goal and subgoal distance on package choice, we plotted the choices on a standard ellipse. Because the ratio of major/minor axis was constant, all house-truck-envelope triplets

could be transformed to a standardized ellipse.[3] This allowed to look at the raw choices. To quantify the degree of influence of each type of distance we conducted a logistic regression on the choice data from phase three. Regressors included (1) the ratio of the distances from the truck to subgoal one and subgoal two, and (2) the ratio of the distances from the truck to the house through subgoal one and subgoal two. To test for significance across subjects, we carried out a two-tailed t test on the population of regression coefficients.

To further characterize the results, we fitted two RL models to each participant's phase-three choice data. One model assigned primary reward only to goal attainment and so was indifferent to subgoal distance *per se*. A second model assigned primary reward to the subgoal as well to the goal.

Value in the first case was the discounted number of steps to the goal, and in the second case it was a sum of discounted number of steps to the subgoal and to the goal. Choice was modeled using a softmax function, including a free inverse temperature parameter. The fmincon function in MATLAB was used to fit discount factor and inverse temperature parameters for both models and reward magnitude for subgoal attainment for the second model. We then compared the fits of the two models calculating Bayes factor for each participant and performing a two-tailed t test on the factors.

Results

The scatter plot on Figure 2.5 shows a clear dissociation of choices based on an ellipse. If subgoal 2 fell within the ellipse, its total distance to be travelled would be smaller than the distance for the reference envelope. The converse would happen if the second envelope was outside of the ellipse. This plot

[3]Because of a technical problem only some ellipses had this ratio, others had a ratio of 5/3. For this reason not all choices could be standardized, as in Figure 2.5. However, this had no impact on the logistic regression.

suggested that the only factor governing choice was the total distance to be travelled passing through the envelope.

The results of the logistic regression confirmed this influence (see Figure 2.6). The average coefficient for goal distance was -7.66 ($SD = 3.5$, $p < .001$), whereas for subgoal distance $-.16$ ($SD = .9$, $p = .43$; see Figure 2.6). All participants, except two, showed large negative coefficients to goal distance. At an individual level, none of coefficients for subgoal distance provided a significant fit. The latter observation held even in a subset of trials where the two delivery options were closely matched in terms of overall distance (with ratios of overall goal distance between .8 and 1.2). The model fits yielded converging results, being that the Bayes factor was 4.31, thus favoring the simpler model with primary reward only at the goal.

Discussion

Participants overwhelmingly preferred subgoals that minimized overall path travelled. This held even in pairs of subgoals that differed little in their goal distance or subgoals that involved an initial travel in the opposite direction to the house. The absence of a significant trend for the subgoal coefficients at the population level strongly mitigates against attaching reward to the attainment of the envelope. The test at the population level could however fail to find individual differences or opposing effects. At the individual level, no fit of the logistic model yielded a significant contribution of subgoal distance. It could still be the case that opposing effects could happen *within* the same participant. If this were the case, we would be able to see a "cloud" of chosen subgoals around the house, when inspecting the raw data points plotted on the transformed ellipse, which we did not observe (Figure 2.5).

Overall these results are consistent with HRL, in that subgoals do not trade off with goals, and goals dominate choice. These findings reassure that any manipulation of the subgoal *does not* yield changes in primary or secondary reward. They do not guarantee, however, that subgoal attain-

ment has something akin to pseudo-reward. This is suggested by the last experiment in this chapter, and by an exclusion of alternative hypotheses for the neural findings in the subsequent chapters.

2.3 Testing Subgoal Approach Behavior

In the previous section there was no observable influence of pseudo-reward on choice behavior, when pitted against reward. However, HRL predicts that when an agent is executing an option, choice should be influenced by option-specific values, driven by pseudo-reward, independently of the top-level value. Moreover, this effect should be clearer when choosing between subgoals of equal goal distance. We would like to ascertain whether this would be the case in two separate experiments (each with a different set of participants). The first experiment offers choice after participants have started to head towards the subgoal. The second experiment also presents choices while the participant is within the option, with the addition that these subgoals have the same overall distance to the goal.

Methods: choice while executing an option

Twenty-two participants (M = 19.4, SD = .87, 15 male) played a video game very similar to the one described before. Task and procedure were the same in all aspects with the exception that choice was offered after participants had started heading towards the subgoal. The initial location of the subgoal was determined to be at least 200 pixels from the house and from the starting location of the truck, and be on an ellipse with major axis twice the minor axis (resolution 1440 x 900 pixels). After the first or the second step, a brief tone was played, the previous envelope disappeared, and two envelopes appeared. One of the envelopes was randomly located on the same ellipse as the initial location, with the constraint of being at a minimal distance of 100 pixels from the house, truck and previous subgoal — this

ellipse was calculated at the moment of the choice. The location of a second envelope was determined randomly to be at least 100 pixels away from the previously calculated icons. After the tone, participants could choose which of the envelopes to pick up in the same way as was described in the previous paradigm, by pressing a key while an envelope was highlighted. The rest of the task proceeded as in the previous experiment. In the second phase, two-thirds of the trials would end after the choice. We used a similar data analysis approach as in the first experiment. After obtaining the coefficients from the logistic regression of subgoal and goal distances, we did a one-tailed t test comparing the subgoal coefficients between the current and the first experiments. Unless otherwise stated, the significance of coefficients for single subjects followed the population trend.

Methods: choice between subgoals of equal overall distance, after start of the option

Nine participants ($M = 20.44$, $SD = .5$, 3 male) played a video game very similar to the one described before. Task and procedure was the same in all aspects with the exception that both envelopes were now on an ellipse (both subgoals are different from the subgoal with which the participant started the task). The location of the subgoals was determined by placing two subgoals randomly on an ellipse with a major/minor axis ratio of 2/1, with a minimal distance of 100 pixels of each other and the other icons.

Results

On both experiments there was no significant increase in influence of the subgoal relative to the first choice experiment (choice while executing an option: mean difference = .13, $p = .71$; choice between equidistant subgoals: mean difference = .13 $p = .66$, Figure 2.7). Results at the individual level were in line with the population trend, in that no participant exhibited a

significant contribution of distance to the subgoal. As in the first experiment, the distance to the goal drove choices of all participants (M = -3.1, $p < .001$).

Discussion

We tested subtler predictions of HRL, whereby people should prefer less action costs to attain a subgoal if they have initiated the option leading to that subgoal, and between subgoals with the same overall actions costs. In HRL, this is dictated by option-specific value functions, reflecting expected discounted pseudo-reward. Contrary to our predictions, neither cognitive manipulation elicited preferences for closer subgoals. In theoretical terms, such pattern of choices is in accordance with a goal-driven reward function. This would not disprove a hierarchical structure of behavior since it is possible to achieve hierarchical control without pseudo-reward (e.g., Parr, 1998).

However, pseudo-reward is widely used for independent reinforcement learning at the level of subtasks (Dietterich, 1998; Sutton et al., 1999), and the evidence that people learn to optimize behavior locally, independently of reward (Diuk et al., 2013) strongly suggests a hierarchical reward function.[4] In addition, the medial frontal response to subtask prediction errors, presented later in this thesis, adds credence to the existence of a separate, subtask reward function. Indirect evidence for pseudo-reward in this task might come from a post-hoc analysis of participants' paths. We tested, for each participant, whether the path from start to choice was indistinguishable from a straight line from start to the pre-choice subgoal location. This was done by regressing the set of {x,y} points, from start to choice, onto a line, then subtracting the slope of the fitted line to the slope of the

[4]We are excluding statistical forms of learning which do not require pseudo-reward. Though they also yield learning at multiple hierarchical levels (e.g., Saffran & Wilson, 2003), they are usually studied in the domain of perception, and are not tied to reinforcement (Turk-Browne & Scholl, 2010).

straight-to-subgoal line, and doing a two tailed t test on the distribution of differences. In this analysis no single participant exhibited a significant difference from the straight-to-subgoal path ($\alpha = .05$). This is interesting because the optimal path for an agent that is not driven by pseudo-reward is a straight line in the direction of the house, and then changing the course depending on the chosen subgoal. Nevertheless, given that the there are only three pre-choice {x,y} points, we cannot safely rely on this analysis.

We can raise several possibilities for the absence of an effect on choices. Firstly, it could be that any scenario where goal distance and subgoal distance are pitted against each other eclipses any preferences for the latter. To our knowledge, there is no precedence for this in the computational literature.[5] On the other hand, competition of information for cognitive processing is a widely acknowledged phenomenon in psychology. In perception, limited capacity leads to processing of only behaviorally relevant stimuli (Desimone & Duncan, 1995). On this line, there could be limited perceptual capacity to evaluate distances simultaneously at a subgoal and goal level. Another possibility, is that goal preferences more automatically control behavior, similarly to the precedence of word reading over color naming (Miller & Cohen, 2001), perhaps driven by an ecological imperative.

However, limited capacity and automaticity do not explain the absence of subgoal preferences when both envelopes are on the ellipse. One possible alternative is that participants terminate the option and return to the root level. In this case, action selection would no longer be governed by option-specific values, even if the action costs to the goal were preserved, but would driven by reward. In the next experiment we sought to have choices happen while performing an option, and to eliminate possible causes for option termination. A possible candidate for eliciting option termination is the imposed choice, and the length of the pause (6 s). Thirty-four participants

[5] If anything, subtask values are given preference over top-level values, something known as recursive optimality (Dietterich, 1998).

chose between pairs of buttons. One button brought the envelope closer and another had the converse effect. Importantly, the change was voluntary and immediate, in other words, participants could elect not to press any of the buttons and the change took place with a minimal amount of pause. In all cases the overall action costs to the goal were respected.

Methods

Thirty-four participants ($M = 20.6$ years, $SD = .92$, 15 male) performed ten deliveries in a first phase. On a second phase, participants could press one of two buttons in the joystick. At the beginning of the experiment participants were told that it would be completely up to them to press the button or not. We avoided using expressions such as "play" with the buttons to discourage an exploratory bias. Participants could press the button at any point until picking up the envelope. One of the buttons would bring the envelope to the point on the ellipse, between the envelope and the truck, that would be closest to 70% of a straight line between truck and envelope, whereas the other button would have the opposite effect, bringing the envelope closer to the house. Both manipulations happened on the ellipse. The change would take place immediately after pressing a button. In case no button was pressed, the trial proceeded as a regular delivery. Participants were not told what the general effect of the buttons was.

There were four blocks of 40 trials, each with a pair of buttons. At the onset of each block participants were told which pair of buttons was available for choice. Buttons were numbered as shown in Figure 2.8. To make sure that participants would not forget which pair was available, we showed the screen informing about the available pair twice. Buttons were paired based on similar ease of access (each red box indicates a pair of buttons Figure 2.8. The effect of each button and the order of presentation of the pairs were randomly assigned for each participant. For each participant and for each block, the share of closer button presses out of all presses was calculated,

and averaged across blocks. A share of .5 meant indifference between the two buttons. A two-tailed t test was then performed on the 34 average shares against the the null hypothesis that the mean share was .5.

Results

On average, .48 of the participants pressed a button in a block, with no difference in pressing rate across blocks ($F(3,132) = .179$, $p = .91$). In spite of this, there was large inter-subject variability ($SD = .48$). 18 out of 34 participants pressed both buttons on all blocks, 7 participants pressed both buttons on 3 blocks and played with only one button type on one block, 6 pressed the two button types on 2 blocks, and only one type on the other 2 blocks, and the remaining 3 participants had 1 or 2 blocks with no press at all. In spite of high press rate, the share of closer button presses was not different from .5 ($M = .48$, $SD = .32$, $p = .75$). In contrast with the previous experiments, the population trend around indifference seemed however to be the result of a mixture of "truly" indifferent and very consistent participants on both preferences. Figure 2.9 illustrates the spectrum of preference: on the extremes one participant pressed the closer button on .98 of the trials, averaged across blocks, and another participant chose the farther button on .78 of the trials.

We conducted a post-hoc logistic regression, in order to quantify the degree of subgoal preference in a manner comparable to the previous experiments. This involved assuming that when the participant pressed a button there were two subgoals of equal goal distance, one farther and another closer to the truck than the subgoal prior to choice. The data of the 4 blocks was treated as a single 120-trial experiment. Only trials with a pressed button were considered. We labeled a random half of the true choices as subgoal 1 and the remaining half as subgoal 2 (the reverse labeling was applied to the counterfactual choices). With these two subgoal labels, the ratio of subgoal distances (distance to subgoal1 / distance to subgoal2) was calculated —

as in the analysis for the previous experiments. We ran a logistic regression with the ratio as predictor and whether the chosen subgoal was subgoal 1 or 2.

The population of coefficients was not significantly different from 0 (M = -.16, SD = 1.5, p = .58). In spite of the mean, 11 participants had a significant contribution of subgoal distance ($p < .05$ using Bonferroni correction; 18 participants with uncorrected $p < .05$, Figure 2.10). These 11 participants had a significantly higher rate of button press, compared with the remaining set ($M = .82$ *vs.* $M = .34$, $p < .001$).

Discussion

As hypothesized, choosing while executing an option, and eliminating the possible trigger of termination (the pause generated by choice) allowed subgoal preferences to be revealed. These were distributed on a spectrum of choice patterns. The results for the subset of participants that preferred a closer envelope are in line with the posited effect of pseudo-reward (black dots in Figure 2.10). An HRL agent should show a preference for earlier/less effortful attainment ot the subgoal. Contrary to our predictions, but in line with the previous experiments, there was a group indifferent to closer subgoals. The subset of indifferent participants might actually be a mixture in itself, given the trending results for 7 participants. In any case, we cannot safely affirm whether these indifferent participants had or had not subgoal preferences. Making the expression of preferences voluntary, and providing no information about the effect of the buttons, might make "true" preferences vulnerable to switch costs, and attention to early learning about the dynamics of the task.

More surprisingly so, we found a group with preferences for farther subgoals. These participants informally mentioned that they liked being closer to the house when picking-up the envelope. In computational terms, this would be equivalent to preferring for higher top-level values at the moment

the option has ended. Diuk et al. (2013) has shown that extended values, dependent on the combination of the values of subtasks, influence behavior and striatal activity at the end of a task and not as information is available. This is suggestive of hierarchical valuation, whereby an agent accesses option-specific values while pursuing a subgoal and only accesses the root reward function, and top-level values, once the subgoal has been attained (in contrast with continuous integration of information).[6] To be clear, an RL agent would not show such results, because the action structure does not reflect subgoal attainment or extended policies, and thus it does not make sense to posit that $V_{s=envelope}$ is given more priority in evaluation that any other state.

2.4 Chapter Discussion

Our data are consistent with an interpretation under which the onset of the choice stimuli triggers a return to the root level, a shift that does not occur in the last experiment, where choices are made without such an exogenous trigger. The latter experiment, importantly, revealed subgoal-related preferences as predicted by HRL. However, the direction and range of these preferences was a surprise, with an interesting pattern of individual diffences. We interpret these in terms of the pseudo-reward function: some people favor low-cost subgoal attainment, others prefer to "set up" for subsequent subtasks. Preferring subgoals that prepares for the initiation set of an ensuing option is the essence of skill chaining, a method of generating options (Konidaris & Barto, 2009). In either case, the data are consistent with the general predictions of HRL, and cannot be explained by flat RL, revealing a scoping or encapsulation of value at different levels of task structure, and revealing a role for such values in learning.

[6] Though some models of *options* with interruption specify that an agent has access to top-level values even during option performance.

Regarding the indifferent participants in the last experiment, it is unlikely that this subset had a flat representation of the task. Hierarchical representations are immediate to humans, even in situations where they are not required or are detrimental (Rosenbaum et al., 1983; Badre et al., 2010; Collins & Frank, 2013). The indifferent participants exhibited a lower rate of button presses, suggesting that their "preferences" are actually the effect of exploration. We posit that, when novelty bonuses were no longer at play, switch costs of pressing a button would eclipse any subtask preference.

In four experiments we examined no attachment of reward to subgoals, and approach of subgoals while performing an option. There are additional predictions from HRL, which we have not focused on. Pseudo-reward should lead to the creation of option policies. Also, after an option policy is learned, we should observe a transfer effect to another task, which can positive or negative, depending on the appropriateness of the option to the task at hand (something we discuss in the last chapter of this thesis).

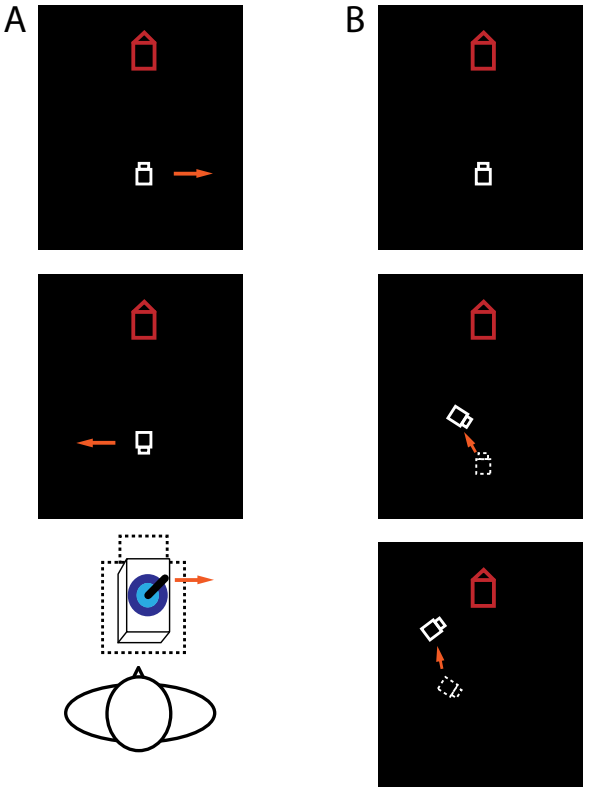

Figure 2.3. Implementation of action costs. (A) Illustration of the effect of a movement command to the right of the joystick — as shown in the figure, actual displacement on the screen depends on orientation of the truck. After every movement command the joystick had to be reset from the outer "Move threshold" (dark blue in the joystick) to the "Restart threshold" (light blue). (B) An example of two movement commands.

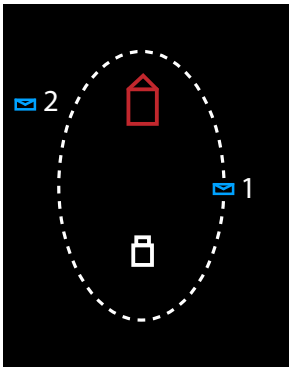

Figure 2.4. Choice between subgoals. The participant would only see the colored elements (the dashed ellipse and the labels would not be shown). Subgoal 2 could be inside or outside of the ellipse.

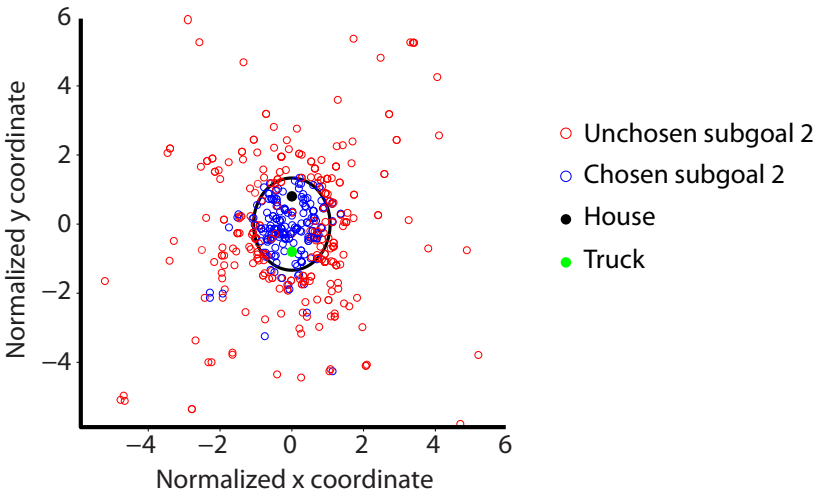

Figure 2.5. Spatial distribution of choices for all participants. This figure represents a minimal and transformed version of the task display. In solid green and black are the truck and house. The solid black line illustrates the position of the reference subgoal (which on each trial could be on any point on the ellipse). The red points indicate a setting where the reference subgoal (on the ellipse), was chosen. The blue points indicate the converse, where the reference subgoal was not chosen. As can be seen, there was a clear demarcation of preferences, based on overall distance to the goal.

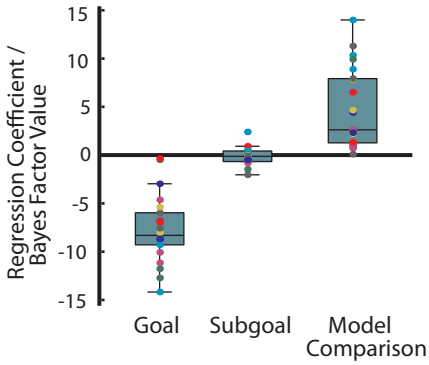

Figure 2.6. Influence of subgoal and goal distances on choice. On the left are the coefficients from a logistic regression with ratios of goal distance, and ratios of subgoal distance. The right box represents the distribution of Bayes Factors, comparing a model with reward at the house, and a model with reward at attainment of the house and the subgoal. The edges of a box are the 25th and 75th percentiles, the line inside a box is the median, and the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually.

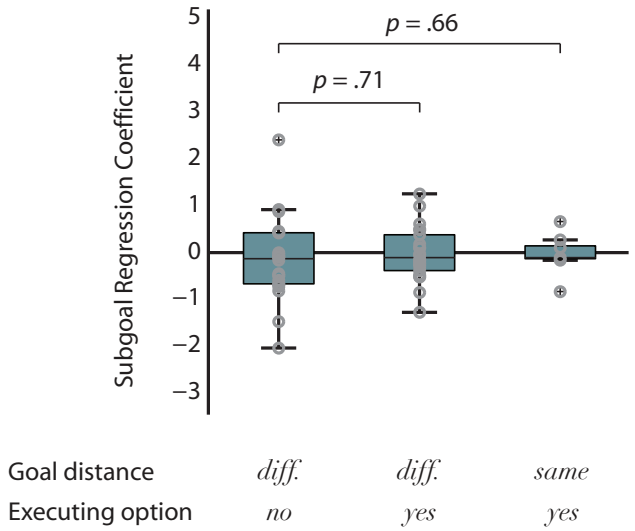

Figure 2.7. Comparison of regression coefficients for subgoal distance for the three experiments. There was no significant increase in subgoal influence for choosing while performing an option or for choice between subgoals of equal distance to the goal. The differences between the experiments are highlighted below the graph (goal distance, choice while executing an option). The edges of a box are the 25th and 75th percentiles, the line inside a box is the median, and the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted with a cross inside a circle.

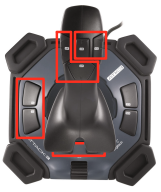

Figure 2.8. Top view of the joystick used for choice. Each block used a different pair of buttons (highlighted by red boxes). One of the buttons was randomly assigned to decrease the costs to the subgoal, whereas the other increased the costs to the subgoal.

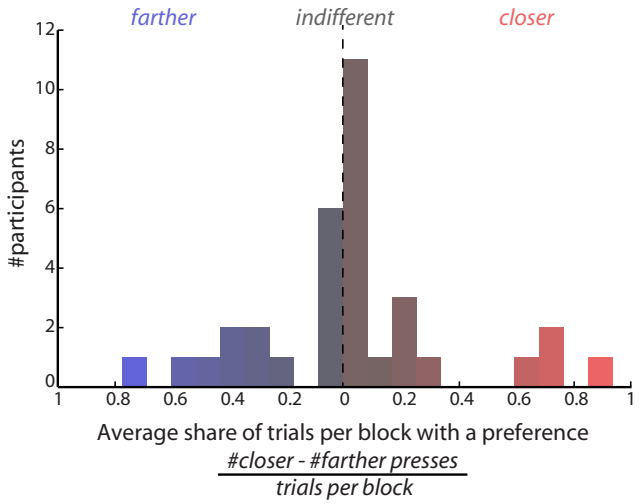

Figure 2.9. Histogram of the share of closer and farther presses. As can be seen there was a spectrum of preferences (the x-axis ranges from pressing the farther button on all trials (blue extreme) to pressing the closer button on all trials (red); in between, the values reflect the share of trials of the difference between the closer and the farther buttons).

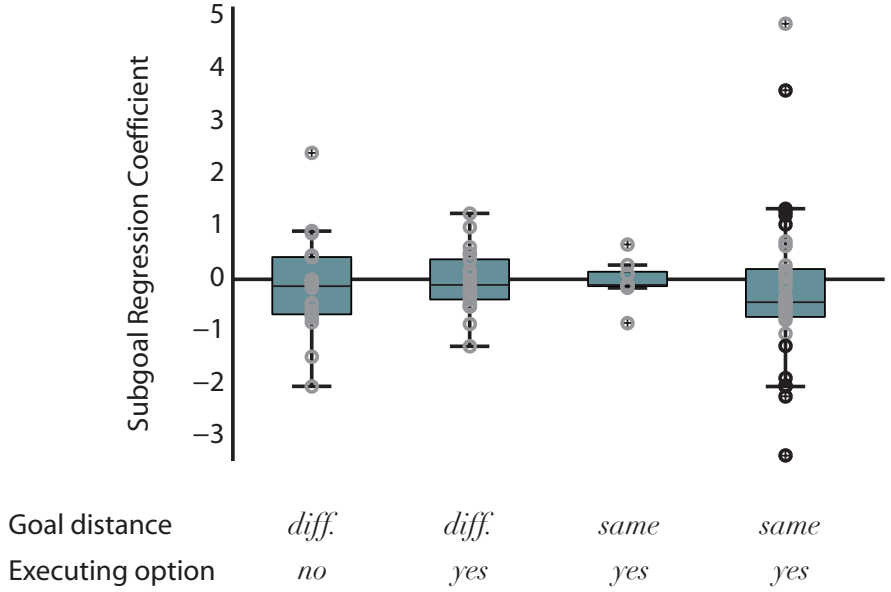

Figure 2.10. Comparison of regression coefficients for subgoal distance for the four experiments. The dots in black are single participants with significant fits for subgoal distance ($p < .05$ corrected — see text for details on the test). The differences between the experiments are highlighted below the graph. The edges of a box are the 25th and 75th percentiles, the line inside a box is the median, and the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted with a cross inside a circle.

Chapter 3

Neural Correlates of Pseudo-Reward Prediction Errors

3.1 Chapter Summary

- In this chapter we describe tests of neural correlates of PPEs, using variants of the task described in the previous chapter. In these paradigms, while the subject is heading towards the subgoal, the subgoal unexpectedly jumps to a new location, which varies in initial distance, but respects overall distance.[1]

- A first EEG experiment sought to examine whether negative PPEs (subgoal jumps to a farther location) would elicit a feedback-related negativity (FRN). This is a potential which reflects anterior cingulate activity, and a known neural correlate of RPEs. We observed a potential to negative PPEs, with an amplitude and location suggestive of the FRN, controlling for errors and conflict.

- In a second study we used fMRI with a similar behavioral paradigm. We found that BOLD signal in dorsal anterior cingulate cortex and anterior insula increased with the magnitude of negative PPEs.

- In a third experiment we examined responses to positive PPEs (subgoal jumps to a closer location). Again we found activity in anterior cingulate and insular cortices correlating with a PPE. More specifically, BOLD signal increased with the magnitude of the positive PPE. No striatal response was observed.

- Overall, these experiments are consistent with a role of anterior cingulate cortex in HRL-related processes, signalling an unsigned prediction error. We found no correlates of signed prediction errors. This is hypothesized to be a result of opposing preferences in the population, as observed in the previous behavioral experiments, yielding an average null response in areas that would respond to PPEs in a signed way.

[1] The first two experiments were published in Ribas-Fernandes, Solway, et al.,(2011).

3.2 Introduction

Learning in HRL occurs at two levels. At a global level, the agent learns to select actions and subroutines so as to efficiently accomplish overall task goals. A fundamental assumption of RL is that goals are defined by their association with reward, and thus the objective at this level is to discover behavior that maximizes long-term cumulative reward. Progress toward this objective is driven by temporal-difference (TD) procedures drawn directly from ordinary RL: following each action or subroutine, a reward-prediction error is generated, indicating whether the behavior yielded an outcome better or worse than initially predicted (see Figure 3.1 and methods section), and this prediction error signal is used to update the behavioral policy. Importantly, outcomes of actions are evaluated with respect to the global goal of maximizing long-term reward.

At a second level, the problem is to learn the subroutines themselves. Intuitively, useful subroutines are designed to accomplish internally-defined subgoals (Singh et al., 2005). For example, in the task of making coffee, one sensible subroutine would aim at adding cream. HRL makes the important assumption that the attainment of such subgoals is associated with a special form of reward, labeled pseudo-reward to distinguish it from external or primary reward. The distinction is critical because subgoals may not themselves be associated with primary reward. For example, adding cream to coffee may bring one closer to that rewarding first sip, but is not itself immediately rewarding. In an HRL context, accomplishment of this subgoal would yield pseudo-reward, but not primary reward. Once the HRL agent enters a subroutine, prediction error signals indicate the degree to which each action has carried the agent toward the currently relevant subgoal and its associated pseudo-reward (see Figure 3.1). Note that these subroutine-specific prediction errors are unique to HRL. In what follows, we refer to

them as pseudo-reward prediction errors (PPE), reserving reward prediction error (RPE) for prediction errors relating to primary reward.

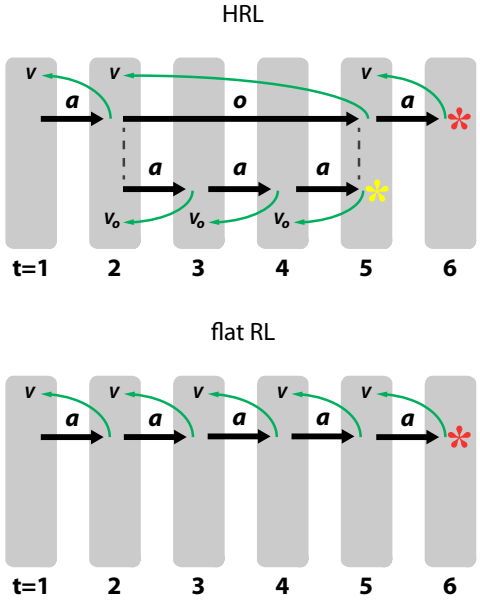

Figure 3.1. Learning updates in HRL and RL. In addition to learning values driven by reward, an HRL agent learns simultaneously at the subtask level. This is driven by pseudo-reward and involves reward-independent updates called pseudo-reward prediction errors (PPE, represented by the lower-facing green arrows) and option-specific values V_o. For comparison, the lower diagram represents the updates in a flat RL agent. Adapted from Botvinick, Niv, and Barto (2009).

In order to make these points concrete, consider the video game illustrated in Figure 3.2, which is based on a benchmark task from the computational HRL literature (Dietterich, 1998). Only the colored elements in the figure appear in the task display. The overall objective of the game is to complete a delivery as quickly as possible, using joystick movements to guide the truck first to the package and from there to the house. It is self-evident

how this task might be represented hierarchically, with delivery serving as the (externally rewarded) top-level goal and acquisition of the package as an obvious subgoal. For an HRL agent, delivery would be associated with primary reward, and acquisition of the package with pseudo-reward. This observation is not meant to suggest that the task must be represented hierarchically. Indeed, it is an established point in the HRL literature that any hierarchical policy has an equivalent non-hierarchical or flat policy (as long as the underlying decision problem satisfies the Markov property). Our neuroimaging experiments proceeded on the assumption that participants would represent the delivery task hierarchically. However, as we discuss later, the neuroimaging results themselves provided convergent evidence for the validity of this assumption.

Consider now a version of the task in which the package sometimes unexpectedly jumps to a new location before the truck reaches it. According to RL, a jump to point A in the figure, or any location within the ellipse shown, should trigger a positive RPE, because the total distance that must be covered in order to deliver the package has decreased. As supported by the behavioral experiments in the previous chapter, we assume temporal/effort discounting, which implies that attaining the goal faster/in less steps is more rewarding. We also assume that current subgoal and goal distances are always immediately known, as they were for our experimental participants from the task display. By the same token, a jump to point B or any other exterior point should trigger a negative RPE. Cases C, D and E are quite different. Here, there is no change in the overall distance to the goal, and so no RPE should be triggered, either in standard RL or in HRL. However, in case C the distance to the subgoal has decreased. According to HRL, a jump to this location should thus trigger a positive PPE. Similarly, a jump to location D should trigger a negative PPE (note that location E is special, being the only location that should trigger neither a RPE nor a PPE). These points are illustrated in Figure 3.2 (right), which

shows RPE and PPE time-courses from simulations of the delivery task based on standard RL and HRL.

To make our computational predictions explicit, we implemented both a standard and a hierarchical RL model of the delivery task, based on the approach laid out in Botvinick, Niv, and Barto (2009). Simulations were performed in Matlab (The Mathworks, Natick, MA); the relevant code is available for download from www.princeton.edu/ matthewb. For the standard RL agent, the state on each step t, labeled s_t, was represented by the goal distance (gd), the distance from the truck to the house, via the package, in units of navigation steps. For the HRL agent, the state was represented by two numbers: gd and the subgoal distance (sd), i.e., the distance between the truck and the package. Goal attainment yielded a reward (r) of one for both agents, and subgoal attainment a pseudo-reward (ρ) of one for the HRL agent. On each step of the task, the agent was assumed to act optimally, that is to take a single step directly toward the package or, later in the task, toward the house. The HRL agent was assumed to select a subroutine (σ) for attaining the package, which also resulted in direct steps toward this subgoal (for details of subtask specification and selection, see Figure 3.1, and Sutton et al., 1999; Botvinick, Niv, & Barto, 2009). For the standard RL agent, the state value at time t, $V(t)$, was defined as γ^{gd} , using a discount factor $\gamma = .9$. The RPE on steps prior to goal attainment was thus:

$$RPE = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) = \gamma^{1+gd_{t+1}} - \gamma^{gd_t} \tag{3.1}$$

The HRL agent calculated RPEs in the same manner, but also calculated PPEs during execution of the subroutine σ . These were based on a subroutine-specific value function (see (Sutton et al., 1999; Botvinick, Niv,

& Barto, 2009)), defined as $V_\sigma(s_t) = \gamma^{sd_t}$. The PPE on each step prior to subgoal attainment was thus:

$$PPE = \rho_{t+1} + \gamma V_\sigma(s_{t+1}) - V_\sigma(s_t) = \gamma^{1+sd_{t+1}} - \gamma^{sd_t} \tag{3.2}$$

To generate the data shown in Figure 3.2, we imposed initial distances (gd, sd) = (949, 524). Following two task steps in the direction of the package, at a point with distances (849, 424), in order to represent jump events distances were changed to (599, 424) for jump type *A*; (1449, 424), type *B*; (849, 124), type *C*; (849, 724), *D*; and (849, 424), *E*. Dashed data series in Figure 3.2 were generated with jumps to (849, 236), *C*; and (849, 574), *D*.

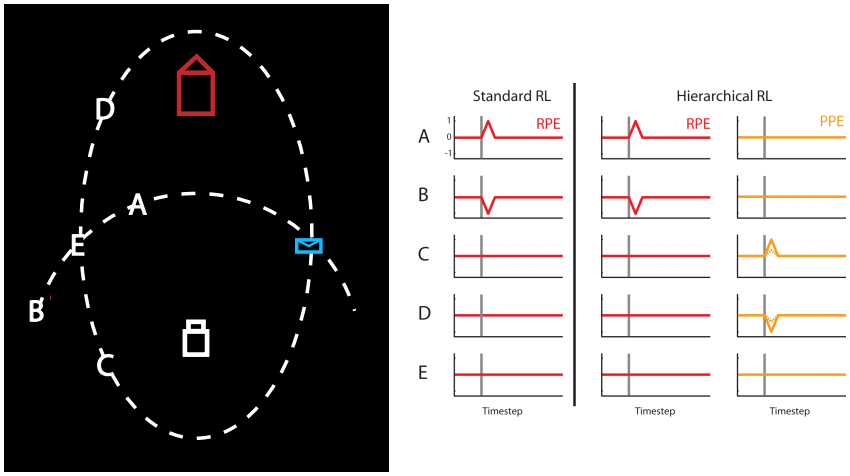

Figure 3.2. Task and predictions from HRL and RL. Left: Task display and underlying geometry of the delivery task. Right: Prediction-error signals generated by standard RL and by HRL in each category of jump event. Grey bars mark the time-step immediately preceding a jump event. Dashed time-courses indicate the PPE generated in *C* and *D* jumps that change the subgoals distance by a smaller amount. For simulation methods, see the methods section below.

These points translate directly into neuroscientific predictions. Previous research has revealed neural correlates of the RPE in numerous structures (Breiter et al., 2001; Holroyd & Coles, 2002; Holroyd et al., 2003; O'Doherty et al., 2003; Ullsperger & von Cramon, 2003; Yacubian et al., 2006; Hare et al., 2008). HRL predicts that neural correlates should also exist for the PPE. To test this, we had neurologically normal participants perform the delivery task from Figure 3.2 while undergoing EEG and, in two further experiments, fMRI.

3.3 An EEG Experiment with Negative PPEs

Motivation

Earlier EEG research indicates that ordinary negative RPEs trigger a midline negativity typically centered on lead Cz, sometimes referred to as the feedback-related negativity or FRN (Miltner, Braun, & Coles, 1997; Holroyd & Coles, 2002; Holroyd et al., 2003). This is thought to originate in the dorsal anterior cingulate cortex (ACC, Gehring and Willoughby, 2002; but see, for an opposing perspective, van Veen, Holroyd, Cohen, Stenger, and Carter, 2004) and to reflect phasic dopaminergic input to this region (Holroyd & Coles, 2002). Based on HRL, we predicted that such frontocentral negativity, suggestive of the FRN, would occur following the critical jumps (type D) in our task.

Methods

Participants. All experimental procedures were approved by the Institutional Review Board of Princeton University. Participants were recruited from the University community and all gave their informed consent. Nine participants were recruited (ages 18-22, $M = 19.7$, 4 males, all right-

handed). All received course credit as compensation, and in addition, received a monetary bonus based on their performance in the task.

Task and procedure. Participants sat at a comfortable distance from a shielded CRT display in a dimly lit, sound attenuating, electrically shielded room. A joystick was held in the right hand (Logitech International, Romanel-sur-Morges, Switzerland). The computerized task was coded using Matlab (The Mathworks, Natick, MA) and the Matlab Psychophysics toolbox, version 3 (Brainard, 1997). On each trial, three display elements appeared: a truck, a package and a house. These objects occupied the vertices of a virtual triangle with vertices at pixel coordinates (0, 180), (150, 30) and (0, -180) relative to the center of the screen (resolution 1024 x 768 pixels), but assuming a random new rotation and reflection at the onset of each trial. The task was to move the truck first to the package and then to the house. Each joystick movement displaced the truck a fixed distance of 50 pixels. As in the behavioral experiments, the orientation of the truck after each step, and participants had to their adapt their responses accordingly. This aspect of the task was intended to assure that intensive spatial processing occurred at each step of the task, rather than only following subgoal displacements. Responses were registered when the joystick was tilted beyond half its maximum displacement. Between responses, the participant was required to restore the joystick to a central position. When the truck passed within 30 pixels of the package, the package moved inside the truck icon and remained there for subsequent moves. When the truck containing the package passed within 35 pixels of the house, the display cleared and a message reading "10c" appeared for a duration of 300 ms (participants were paid their cumulative earnings at the end of the experiment). A central fixation cross then appeared for 700 ms before the onset of the next trial. On every trial, after the first, second or third truck movement, a brief tone occurred and the package flashed for an interval of 200 ms, during which any joystick inputs were ignored. On one third of such occasions, the pack-

age remained in its original location. On the remaining trials, at the onset of the tone, the package jumped to a new location. In half of such cases, the distance between the packages new position and the truck position was unchanged by the jump (case E in Figure 3.2). In the remaining cases, the distance from the truck to the package was increased by the jump, although the total distance from the truck to the house (via the package) remained the same (case D in the figure). In these cases, the jump always carried the package across an imaginary line connecting the truck and the house, and always resulted in a package-to-house distance of 160 pixels. In all three conditions the package would be on an ellipse defined by the locations of the old subgoal, the house and the position of the truck at the time of the jump. By the definition of an ellipse overall distance to the house was preserved. At the outset of the experiment, each participant completed a fifteen minute training session, which was followed by the hour-long EEG testing session. Participants completed 190 trials on average (range 128-231). Trials were grouped into blocks, each containing six trials: two trials in which the position of the package did not change, two involving type E jumps and two type D jumps. The order in which trials of a particular type occurred was pseudorandom within a block. Participants were given an opportunity to rest for a brief period between task blocks.

Data acquisition. EEG data were recorded using Neuroscan (Charlotte, NC) caps with 128 electrodes and a Sensorium (Charlotte, VT) EPA-6 amplifier. The signal was sampled at 1000 Hz. All data were referenced online to a chin electrode, and after excluding bad channels were rereferenced to the average signal across all remaining channels (Hestvik, Maxfield, Schwartz, & Shafer, 2007). EOG data were recorded using a single electrode placed below the left eye. Ocular artifacts were detected by thresholding a slow moving average of the activity in this channel, and trials with artifacts were not included in the analysis. Less than four trials per subject

matched this criterion and were excluded from the analysis (less than two per condition).

Data analysis. Epochs of 1000 ms (200 ms baseline) were extracted from each trial, time-locked to the package's change in position. The mean level of activity during the baseline interval was subtracted from each epoch. Trials containing type D jump were separated from trials containing jumps of type E, and ERPs were computed for each condition and participant by averaging the corresponding epochs. The ERPs shown in Figure 3.3 were computed by averaging across participants. The PPE effect was quantified in electrode Cz, following Holroyd and Coles (2002). The PPE effect was quantified for each subject by taking the mean voltage during the time window from 200 to 600 ms following each jump, for the two jump types. A one-tailed paired t test was used to evaluate the hypothesis that type D jumps elicited a more negative potential than type E jumps. For comparability with previous studies, topographic plots are shown for electrodes FP1, FP2, AFz, F3, Fz, F4, FT7, FC3, FCz, FC4, FT8, T7, C3, Cz, C4, T8, TP7, CP3, CPz, CP4, TP8, P7, P3, Pz, P4, P8, O1, Oz, O2 (as in Yeung, Holroyd, & Cohen, 2005; F7 and F8 were an exception, given that the used cap did not have these electrode locations).

Results

The EEG experiment included nine participants, who performed the delivery task for a total of 60 minutes (190 delivery trials on average per participant). One third of trials involved a jump event of type D from Figure 3.2; these events were intended to elicit a negative PPE. Stimulus-aligned EEG averages indicated that class-D jump events triggered a phasic negativity in the EEG ($p < .01$ at Cz; Figure 3.3, left), relative to the E-jump control

condition.[2] This negativity was largest in the fronto-central midline leads (including electrode Cz, see Figure 3.3, right).

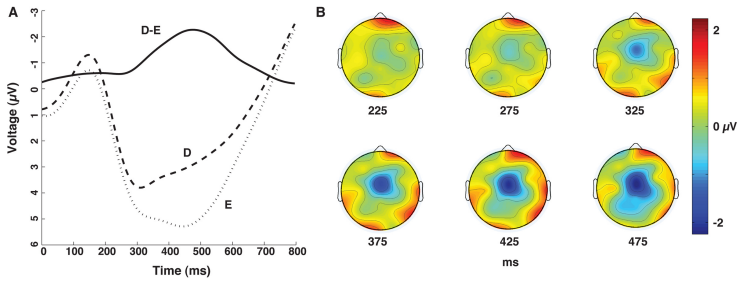

Figure 3.3. Evoked responses at the moment of jump. (A) Evoked potentials at electrode Cz, aligned to jump events, averaged across participants. *D* and *E* refer to jump destinations in Figure 3.2. The data-series labeled *D* - *E* shows the difference between curves *D* and *E*, isolating the PPE effect. (B) Scalp topography for condition *D*, with baseline condition *E* subtracted (topography plotted on the same grid used in Yeung et al., 2005).

Controlling for response conflict, errors and shifts of attention. It was important to evaluate whether the ERP effect observed might reflect error or response conflict detection, factors that have been shown in previous studies to induce phasic midline negativities (Botvinick, Nystrom, Fissell, Carter, & Cohen, 1999; Yeung, Botvinick, & Cohen, 2004; Krigolson & Holroyd, 2006). To rule out an explanation in terms of error-detection, we conducted an analysis that excluded trials where errors occurred. Although there is no discrete criterion for response corrections in the task, it is possible to distinguish between highly accurate and less accurate responses. We defined response accuracy in terms of the angle between the perfect joystick movement (the movement that would have taken the truck directly toward the package) and the actual movement, setting an upper bound of 22.5 °

[2]Like the ERP obtained in this study, the FRN sometimes takes the form of a relative negativity occupying the positive voltage domain, rather than absolute negativity (for germane examples, see Nieuwenhuis et al., 2005; Yeung et al., 2005).

for highly accurate responses, based on an inspection of the response distribution (Figure 3.4A). For clarity, we also only considered trials where the package displacement required a change in the truck path spanning at least 45 °. Repeating our original ERP analysis, focusing only on trials involving highly accurate responses, yielded the ERP data shown in Figure 3.4B. As in the original analysis, the difference between jumps of type D and E was significant ($p = .019$).

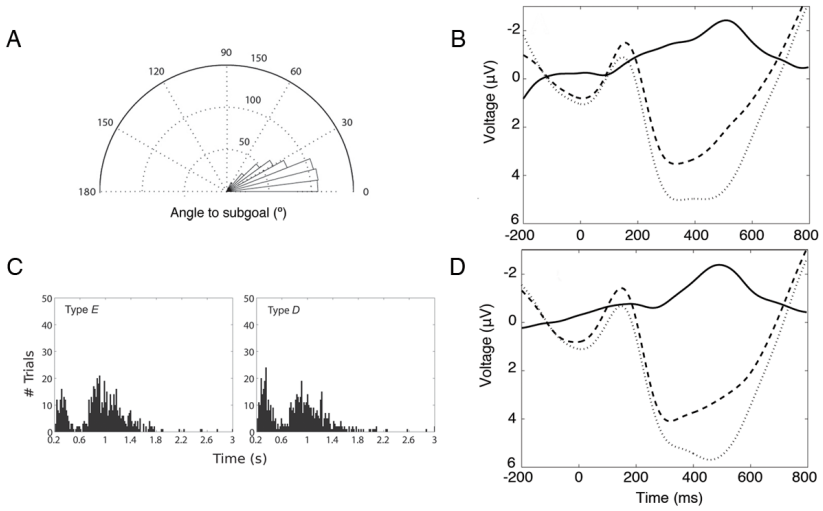

Figure 3.4. Accuracy, reaction times and evoked potentials conditioned on these variables. (A) Polar accuracy plot for the movement command before the subgoal jump. 0 ° is a perfect movement in the direction of the subgoal. Left and right commands are shown collapsed. (B) Evoked potentials at electrode Cz, aligned to jump events and difference wave, conditioned on highly accurate responses. Dashed line corresponds to class D events, grey solid line to E events and the black solid to the difference D - E. (C) Reaction time distributions for type E and D jumps. (D) Evoked potentials at electrode Cz, aligned to jump events and difference wave, conditioned on slow responses. Dashed line corresponds to class D events, grey solid line to E events and the black solid to the difference D - E.

The other alternative explanation we wished to evaluate was related to conflict detection. It was possible that type D jumps caused greater response conflict than type E, perhaps because a greater time was needed to pin down the direction to the new package location (more distant in case D than E). In order to test this explanation, we adopted the common approach of treating reaction time (RT) as an index of conflict. Considering only data from trials with highly accurate responses, mean RT in condition D (1013 ms) did not differ significantly from mean RT for condition E (M = 1049 ms, paired two-tailed t-test, $p = .39$). In fact, unconditioned on accuracy, mean RT following type D jumps (849 ms) was smaller than that following type E jumps (926 ms, paired two-tailed t test, $p < .01$), further militating against an explanation based on conflict. RT distributions for responses immediately following type D and E jumps (collapsing across participants) are shown in Figure 3.4C. RTs in both conditions displayed a clear bimodal distribution, and the difference in mean RT could be largely attributed to a difference in the proportion of fast (and relatively inaccurate) responses versus that of slower (and more accurate) responses. To control for RT, we limited consideration to responses that fell within the slower component of the bimodal distribution in both conditions. The mean RT within the resulting samples (1077 ms for type D, 1075 ms for type E) did not differ significantly across the two conditions (paired two-tailed t test, $p = .98$), nor did the proportion of inaccurate responses, as defined earlier (49.98% for type D vs. 55.10% for type E, paired two-tailed t test $p = .08$). An ERP analysis focusing on this matched data subset of slow responses yielded a robust PPE effect ($p = .02$, Figure 3.4D). EEG correlates of shifts of attention. Figure 3.5 shows the electrode potential at Cz for conditions involving a shift of attention (average of conditions D and E) and the no jump condition.

Note that, in previous EEG research, exogenous shifts of attention have been associated with a midline positivity, the amplitude of which grows with

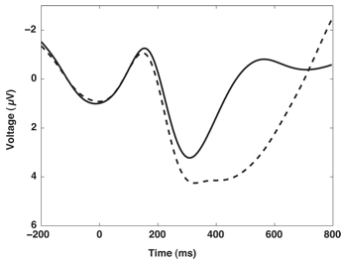

Figure 3.5. ERP for conditions involving a shift of attention (E and D; dashed line) and the condition with no jump (solid line) in electrode Cz. 0 ms is the moment when the package flashes yellow and a tone is played.

stimulus eccentricity (Yamaguchi, Tsuchiya, & Kobayashi, 1995). A midline negativity has been reported in at least one study focusing on endogenous attention (Grent-'t Jong & Woldorff, 2007), but the timing of this potential differed dramatically from the difference wave in our EEG study, peaking at 1000-1200 milliseconds post-stimulus, hundreds of milliseconds after our effect ended. In fact, we observed such a positivity in our own data, in Cz, when we compared jump events (D and E) against occasions where the subgoal stayed put, an analysis specifically designed to uncover attentional effects (see Figure 3.5). In contrast, the PPE effect in our data took the form of a negative difference wave (see Figure 3.3), consistent with the predictions of HRL and contrary to those proceeding from previous research on attention.

Discussion

Like the FRN, we observed a fronto-central negativity to negative PPEs. Although the observed negativity peaked later than the typical FRN, its timing is consistent with studies of equivalent complexity of feedback (Baker & Holroyd, 2011). As mentioned before, fronto-central negativities around 200 ms can reflect negative RPEs (Miltner et al., 1997; Gehring & Willoughby, 2002; Holroyd & Coles, 2002), but also errors (Gehring, Goss, Coles, Meyer,

& Donchin, 1993; Krigolson & Holroyd, 2006) or response conflict (Yeung et al., 2004). Post-hoc analyses of EEG data based on RTs and accuracy showed that the observed negativity was independent of the variables, thus suggesting that it was indeed a response to a prediction error. Overall, this is suggestive of the involvement of ACC in coding negative PPEs, and perhaps mesocortical dopamine. In the next experiment we repeat the paradigm, eliciting negative PPEs, using fMRI.

3.4 An fMRI Study of Negative PPEs

Methods

Participants. Participants were recruited from the University community and all gave their informed consent. For the first fMRI experiment, 33 participants were recruited (ages 18-37, $M = 21.2$, 20 males, all right-handed). Three participants were excluded: two because of technical problems and one who was unable to complete the task in the available time. All participants received monetary compensation at a departmental standard rate.

Task and procedure. An MR compatible joystick (MagConcept, Redwood City, CA) was used. The task was identical to the one used in the EEG experiment, with the following exceptions. Initial positions of the icons were randomly assigned to the screen respecting a minimal distance of 150 pixels between icons. On type D jumps, the destination of the package was chosen randomly from all locations satisfying the conditions that they (1) increase truck-to-package distance, but (2) leave total path length to the goal (house) unchanged. The forced delay involved in the task interruption (tone, package flashing) totaled 900 ms. At the completion of each delivery, the message "Congratulations!" was displayed for 1000 ms, followed by a fixation cross that remained on screen for 6000 ms. The first fMRI experiment consisted of three parts: a fifteen minute behavioral practice outside the scanner, an eight minute practice inside the scanner during

structural scan acquisition and a third phase of approximately forty-five minutes, where functional data were collected. During functional scanning, 90 trials were completed, in six runs of fifteen trials each. At the beginning and end of each run a central fixation cross was displayed for 10000 ms. The average run length was 7.5 minutes (range 5.7-11).

Image acquisition. Data were acquired with a 3 T Siemens Allegra (Malvern, PA) head-only MRI scanner, with a circularly polarized head volume coil. High-resolution (1 mm^3 voxels) T1-weighted structural images were acquired with an MP-RAGE pulse sequence at the beginning of the scanning session. Functional data were acquired using a high-resolution echo-planar imaging pulse sequence (3 x 3 x 3 mm voxels, 34 contiguous slices, 3 mm thick, interleaved acquisition, TR of 2000 ms, TE of 30 ms, flip angle 90 °, field of view 192 mm, aligned with the Anterior Commissure - Posterior Commissure plane). The first five volumes of each run were ignored.

Data analysis. Data were analyzed using AFNI software (Cox, 1996). The T1-weighted anatomical images were aligned to the functional data. Functional data was corrected for interleaved acquisition using Fourier interpolation. Head motion parameters were estimated and corrected allowing six-parameter rigid body transformations, referenced to the initial image of the first functional run. A whole-brain mask for each participant was created using the union of a mask for the first and last functional images. Spikes in the data were removed and replaced with an interpolated data point. Data was spatially smoothed until spatial autocorrelation was approximated by a 6 mm FHWM Gaussian kernel. Each voxels signal was converted to percent change by normalizing it based on intensity. The mean image for each volume was calculated and used later as baseline regressor in the general linear model, except in the region of interest analysis where the mean image of the whole brain was not subtracted from the data. Anatomical images were used to estimate normalization parameters to a template in Talairach space

(Talairach & Tournoux, 1988), using SPM5 (www.fil.ion.ucl.ac.uk/spm/). These transformations were applied to parameter estimates from the general linear model.

General linear model analysis. For each participant we created a design matrix modeling experimental events and including events of no interest. At the time of an experimental event we defined an impulse and convolved it with a hemodynamic response. The following regressors were included in the model: (a) an indicator variable marking the occurrence of all auditory tone / package flash events, (b) an indicator variable marking the occurrence of all jump events (spanning jump types E and D), (c) an indicator variable marking the occurrence of type D jumps, (d) a parametric regressor indicating the change in distance to subgoal induced by each D jumps, mean-centered, (e and f) indicator variables marking subgoal and goal attainment, and (g) an indicator variable marking all periods of task performance, from the initial presentation of the icons to the end of the trial. Also included were head motion parameters, and first to third order polynomial regressors to regress out scanner drift effects. A global signal regressor was also included (comparable analyses omitting the global signal regressor yielded statistically significant PPE effects in ACC, bilateral insula and lingual gyrus, in locations highly overlapping with those reported subtracting global signal).

Group analysis. For each regressor and for each voxel we tested the sample of 30 subject-specific coefficients against zero in a two-tailed t test. We defined a threshold of $p = .01$ and applied correction for multiple comparison based on cluster size, using Monte Carlo simulations as implemented in AFNIs AlphaSim. We report results at a corrected $p < .01$.

Follow-up analysis. Our experimental prediction related to the change in distance between truck and package induced by type-D jump events, i.e., the change in distance to subgoal, or PPE effect. However, jump events also varied in the degree to which they displaced the package (i.e., the distance

from its original position to its post-jump position), and this distance correlated moderately with the increase in subgoal distance. It was therefore necessary to evaluate whether the regions of activation identified in our primary GLM analysis might simply be responding to subgoal displacement (and possible attendant visuospatial or motor processes), rather than the increase in distance to subgoal. To this end, we looked at each area identified in the primary GLM, asking whether the area continued to a show significant PPE effect even after this regressor was made orthogonal to subgoal displacement. In order to avoid bias in this procedure, we employed a leave-one-out cross-validation approach, as follows. For every sub-group of 29 participants (from the total sample of 30) we re-ran the original GLM, identifying voxels that (1) showed the PPE effect at significance threshold of $p = .05$ (cluster-size thresholded to compensate for multiple comparisons), and (2) fell within 33 mm of the peak-activation coordinates for one of the six clusters identified in our primary GLM (dorsal anterior cingulate, bilateral anterior insulae, left lingual gyrus, left inferior frontal gyrus, and right supramarginal gyrus). The resulting clusters were used as regions of interest (ROI) for the critical test. Focusing on the one subject omitted from each 29-subject sub-sample, we calculated the mean coefficient within each ROI for the PPE effect, after orthogonalizing the PPE regressor to subgoal displacement (and including subgoal displacement in the GLM). This yielded thirty coefficients per ROI. Each set was tested for difference from zero, using a two-twailed t test.

Region of interest analysis. We defined nucleus accumbens (NAcc) based on anatomical boundaries on a high-resolution T1-weighted image for each participant; habenula, using peak Talairach coordinates (5, 25, 8), guided by Ullsperger and von Cramon (2003), surrounded by a sphere with a radius of 6 mm (Salas & Montague, 2010); and amygdala, drawn using the Talairach atlas in AFNI. Mean coefficients were extracted from these regions for each participant. Reported coefficients for all regions of interest

are from general linear model analyses without subtraction of global signal. The sample of 30 subject-specific coefficients were tested against zero in a two-tailed t test, with a threshold of $p < .05$.[3]

Results

A group of thirty participants performed a slightly different version of the delivery task, again designed to elicit negative PPEs. As in the EEG experiment, one-third of trials included a jump of type *D* (as in Figure 3.2) and another third included a jump of type *E*. Type *D* jumps, by increasing the distance to the subgoal, were again intended to trigger a PPE. However, in the fMRI version of the task, unlike the EEG version, the exact increase in subgoal distance varied across trials. Type *D* jumps were therefore intended to induce PPEs that varied in magnitude (see Figure 3.2). Our analyses took a model-based approach (O'Doherty, Hampton, & Kim, 2007), testing for regions that showed phasic activation correlating positively with predicted PPE size.

A whole-brain general linear model analysis, thresholded at $p < .01$ (cluster-size thresholded to correct for multiple comparisons), revealed such a correlation in the dorsal anterior cingulate cortex (ACC; Figure 3.6, case *D*). This region has been proposed to contain the generator of the FRN (Holroyd and Coles, 2002, although see Niewenhuis et al., 2005). In this regard, the fMRI result is consistent with the result of our EEG experiment. The same parametric fMRI effect was also observed bilaterally in the anterior insula, a region often coactivated with ACC in the setting of unanticipated negative events (Phan, Wager, Taylor, & I, 2004). The effect was also detected in right supramarginal gyrus, the medial part of lingual gyrus, and, with a negative coefficient, in the left inferior frontal gyrus. However, in a

[3]These analyses were intended to bring greater statistical power to bear on these regions, in part because their small size may have undermined our ability to detect activation in them in our whole-brain analysis, where a cluster-size threshold was employed.

follow-up analysis we controlled for subgoal displacement (e.g., the distance between the original package location and point D in Figure 3.2), a nuisance variable moderately correlated, across trials, with the change in distance to subgoal. Within this analysis, only ACC ($p < .01$), bilateral anterior insula ($p < .01$ left, $p < .05$ right) and right lingual gyrus ($p < .01$) continued to show significant correlations with the PPE. In the series of region-of-interest (ROI) analyses, the habenular complex was found to display greater activity following type D than type E jumps ($p < .05$), consistent with the idea that this structure is also engaged by negative PPEs. A comparable effect was also observed in the right, though not left, amygdala ($p < .05$). In the nucleus accumbens (NAcc) no significant PPE effect was observed (tests for average bilateral accumbens: $p = .23$ for parametric PPE, .09 for categorical PPE, results were comparable on left and right accumbens, and with the inclusion of ventral caudate and ventral putamen).

Discussion

This experiment yielded a significant parametric PPE effect in several regions. One additional aspect of the results that deserves comment is the fact that these same regions did not display a statistically significant categorical effect. That is, while their activation scaled with the magnitude of the subgoal-distance increase induced by type D jumps, the mean activation induced by type D jumps was not significantly greater than that induced by type E jumps. Two possible explanations can be offered for this aspect of the results. First, it should be noted that the average increase in subgoal distance across all trials in the experiment was well above zero. Taking this into account, on a precise HRL account, type E jumps should in fact have induced a small positive PPE. For simplicity, in deriving our experimental predictions, we assumed that the PPE was calculated against a reference or expected subgoal-distance change of zero. This difference between the assumptions of our model and a strict HRL account may at least partially

account for the details of our GLM results. On a more prosaic level, it should be noted that, across trials, the increase in subgoal distance was heavily skewed to the right. This may have undermined power for detecting a mean effect of jump type, making it easier to detect the parametric effect that we in fact obtained in the main GLM analyses. Further experimentation is called for to evaluate the merit of these two interpretations. As noted in the introduction, the design of our neuroimaging experiments reflected a presumption that participants would represent and perform the delivery task in a hierarchical manner. However, as also intimated in the introduction of this chapter, we also view our experimental results as providing evidence supporting that assumption. Specifically, our behavioural study provided evidence against a non-hierarchical or flat RL account involving primary reward at subgoal attainment, and the EEG and fMRI results could not be easily explained by a flat RL account with no reward at subgoal.

In the nucleus accumbens (NAcc), where some studies have observed deactivation accompanying negative RPEs (Knutson, Taylor, Kaufman, & Peterson, 2005), no significant PPE effect was observed. However, it should be noted that NAcc deactivation with negative RPEs has been an inconsistent finding in previous work (see, e.g., O'Doherty, Buchanan, Seymour, & Dolan, 2006; Cooper & Knutson, 2008). More robust is the association between NAcc activation and positive RPEs (Seymour et al., 2004; Hare et al., 2008; Niv, 2009).

We predicted, based on HRL, that neural structures previously proposed to encode temporal-difference RPEs should also respond to PPEs. Negative PPEs were found to engage three structures previously reported to show activation with negative RPEs: ACC, amygdala and habenula. On a cautionary note, findings purported to originate in the habenular complex may be due to spatial spread of signal from other structures, and detection of habenular activity using fMRI might require methods with finer spatial resolution (Lawson, Drevets, & Roiser, 2012). Of course, the association

of these neural responses with the relevant task events does not uniquely support an interpretation in terms of HRL (see Poldrack, 2006). However, aspects of either the task or the experimental results do militate against the most tempting alternative interpretations. Our precursory behavioral studies provided evidence against primary reward at subgoal attainment, closing off an interpretation of the neuroimaging data in terms of standard RL. Given previous findings pertaining to the ACC, the effect we observed in this structure might be conjectured to reflect response conflict or error detection (Botvinick et al., 1999; Yeung et al., 2004; Krigolson & Holroyd, 2006). However, additional analyses of the EEG data indicated that the PPE effect persisted even after controlling for response accuracy and for response latency, each commonly regarded as an index of response conflict). Another alternative that must be addressed relates to spatial attention. Jump events in our neuroimaging experiments presumably triggered shifts in attention, often complete with eye movements, and it is important to consider the possibility that differences between conditions on this level may have contributed to our central findings. While further experiments may be useful in pinning down the precise role of attention in our task, there are several aspects of the present results that argue against an interpretation based purely on attention. Our fMRI results also resist an interpretation based on spatial attention alone. We did find activation in or near the frontal eye fields and in the superior parietal cortex regions classically associated with shifts of attention (Corbetta, Patel, & Shulman, 2008) in an analysis contrasting all jump events with trials where the subgoal remained in its original location (Figure 3.6, jump to E). However, as reported above, activity in these regions did not show any significant correlation with our PPE regressor (see Figure 3.6, jump to D).

If one does adopt an HRL-based interpretation of the present results, then several interesting questions follow. Given the prevailing view that temporal-difference RPEs are signaled by phasic changes in dopaminergic

activity (Schultz et al., 1997), one obvious question is whether the PPE might be signaled via the same channel. ACC activity in association with negative RPEs has been proposed to reflect phasic reductions in dopaminergic input (Holroyd & Coles, 2002), and the habenula has been proposed to provide suppressive input to midbrain dopaminergic nuclei (Christoph, Leonzio, & Wilcox, 1986; Matsumoto & Hikosaka, 2007). The implication of ACC and habenula in the present study thus provide tentative, indirect support for dopaminergic involvement in HRL. At the same time, it should be noted that some ambiguity surrounds the role of dopamine in driving reward-outcome responses, particularly within the ACC (for a detailed review, see Jocham & Ullsperger, 2009). The present findings must thus be interpreted with appropriate circumspection. Again, it should be noted that our HRL-based interpretation does not necessarily require a role for dopamine in generating the observed neural events.

3.5 An fMRI Study of Positive PPEs

Introduction

In this section we now examine the converse case, where costs for subgoal attainment are suddenly decreased. As illustrated in Figure 3.2, this triggers a positive PPE. The association between ventral striatal activity and positive prediction errors is stronger than with negative prediction errors (Seymour et al., 2004; Hare et al., 2008; Niv, 2009). Therefore, this paradigm is a better testbed for the possible association between ventral striatum and PPE. This of couse assumes that in fact what is being triggered by the jump to location C is a positive PE. The last behavioral experiment shows that some participants may show a preference for farther subgoals. In any case, it is not a neutral event and PE should be triggered. Also, in this experiment we decided to aim for a strong main effect instead of a parametric effect.

For that reason all PE jumps head towards a similar location (see methods section to why this cannot be set prior to the trial).

Methods

Participants. Participants were recruited from the University community and all gave their informed consent. 30 participants were recruited (ages 18-25, $M = 20.5$, 11 males, all were right-handed). All participants received monetary compensation at a departmental standard rate. In order to further encourage performance, participants also received a small monetary bonus based on task performance.

Task and procedure. An MR compatible joystick (MagConcept, Redwood City, CA) was used. The initial positions of the icons were rotations or reflections, varied randomly, of a pre-established arrangement of icons of a predetermined triangle with vertices truck (0, 200), package (151, -165) and house (0, -200) (coordinates are in pixels, referenced to the center of the screen, 1024 x 768 pixels).

On every trial, after the first, second or third truck movement, a brief tone occurred and the package flashed for an interval of 900 ms, during which any joystick inputs were ignored. On one third of such occasions, the package remained in its original location. On the remaining trials, at the onset of the tone, the package jumped to a new location. In half of such cases, the distance between the packages new position and the truck position was unchanged by the jump (case E in Figure 3.2). On the remaining third, a type C jump would happen, the destination of the package was chosen such that (1) the distance between truck and package always decreased to 120 pixels and (2) leave total path length to the goal (house) unchanged. At the completion of each delivery, the message 10 appeared for 500 ms, indicating the bonus earned for that trial. Immediately following this, a fixation cross appeared for 2500 ms, followed by onset of the next trial. The experiment consisted of three parts: a fifteen minute behavioral practice outside the

scanner, an eight minute practice inside the scanner during structural scan acquisition and a third phase of approximately forty-five minutes, where functional data were collected. During functional scanning, 90 trials were completed, in six runs of fifteen trials each. At the beginning and end of each run a central fixation cross was displayed for 10000 ms. The average run length was 6.8 minutes (range 4.7-10.7).

Image acquisition. Image acquisition protocols was the same as in the first fMRI experiment.

Data analysis. The procedure for preprocessing data was similar to the one used in previous experiment.

General linear model analysis. For each participant we created a design matrix modeling experimental events and including events of no interest. At the time of an experimental event we defined an impulse and convolved it with a hemodynamic response. The following regressors were included in the model: (a) an indicator variable marking the occurrence of all auditory tone / package flash events, (b) an indicator variable marking the occurrence of jump types E and C, (c) an indicator variable marking the occurrence of type C jumps, (d) a parametric regressor indicating the change in distance to subgoal induced by each or C jumps, mean-centered, (e and f) indicator variables marking subgoal and goal attainment, and (g) an indicator variable marking all periods of task performance, from the initial presentation of the icons to the end of the trial. Also included were head motion parameters, and first to third order polynomial regressors to regress out scanner drift effects.

Group analysis. For each regressor and for each voxel we tested the sample of 30 subject-specific coefficients against zero in a two-tailed t test. We defined a threshold of $p = .01$ and applied correction for multiple comparison based on cluster size, using Monte Carlo simulations as implemented in AFNIs AlphaSim. We report results at a corrected $p < .01$.

Region of interest analysis. We defined nucleus accumbens (NAcc) based on anatomical boundaries on a high-resolution T1-weighted image for each participant; habenula, using peak Talairach coordinates (5, 25, 8), guided by Ullsperger and von Cramon (2003), surrounded by a sphere with a radius of 6 mm (Salas & Montague, 2010); and amygdala, drawn using the Talairach atlas in AFNI. Mean coefficients were extracted from these regions for each participant. Reported coefficients for all regions of interest are from general linear model analyses without subtraction of global signal. The sample of 30 subject-specific coefficients were tested against zero in a two-tailed t test, with a threshold of $p < .05$.

Results

At a whole brain level, in a surprising result, an increase BOLD to C jumps relative to jumps E was observed in dorsal anterior cingulate and bilateral anterior insula ($p < .05$ corrected) (Figure 3.6, jump to C). Another region that survived correction was lingual gyrus in a comparable location to the one observed in the previous study. There was no significant response to the variation in subgoal distance. The control regressors $E+C$ and tone/flash/forced delay showed a similar pattern to the same contrast in the previous fMRI experiment, of frontal eye fields and superior parietal cortex (see Figure 3.6). The ROI analysis yielded no significant response in bilateral NAcc ($p = .94$, and qualitatively the same result for ventral striatum), habenula ($p = .52$) or amygdala ($p = .14$). Results were comparable results for unilateral tests. We discuss these results in the next section, together with the findings from the previous studies.

3.6 Chapter Discussion

Dorsal anterior cingulate cortex is known to respond to reward prediction errors (Holroyd et al.,2004, and for a general review Rushworth, Noonan,

Boorman, Walton, and Behrens, 2011). RPEs, however, are far from being the only eliciting stimulus of dACC. It is also known to respond to conflict, errors in performance. We can exclude conflict or error detection given that reaction times are not higher for correct trials, nor are error rates between for condition *C*, compared with *E*. Though the dACC is not part of the canonical set of areas that responds to stimulus-driven attention, such as the temporo-parietal junction, inferior parietal sulcus and right middle frontal gyrus or frontal eye fields (Corbetta et al., 2008), it can be asked whether the response we observed is driven by shear visual displacement, triggering a shift of attention. The experimental condition, *C*, has less distance or angle of visual displacement than the control condition, *E*. Consistently, these areas responded to the occurence of a jump, *C* and was observed in *E*, rather than *C* in isolation. By exclusion, we can say that the response is a prediction error. However, we cannot for certainty say whether it is a positive or negative PE, particularly given the finding that some people prefer subgoal locations to be closer, and others farther, independently of overall distance, though not both.

The dACC has been reported to respond to both positive and negative PEs with increases in activity, both in BOLD and in single-unit measurements (Hayden, Heilbronner, Pearson, & Platt, 2011; Roesch, Esber, Li, Daw, & Schoenbaum, 2012, for a review), in contrast with earlier findings (Holroyd et al., 2004). This type of response is consistent with a learning model based on surprise (Pearce & Hall, 1980; Pearce, Kaye, & Hall, 1982). This means that the direction of the effect we observed is not telling of the valence of the PE. In the behavioral study, we reported a spectrum of preferences for closer to farther subgoal. It might be possible that for a subgroup of participants the change elicited a positive PE, for others a negative PE and for others nothing at all. Crucially none of the PEs is associated with any change with reward delivery and none of these changes would happen if people were not representing the task hierarchically.

In contrast with the dACC, we did not observe any effect in canonical signed RPE areas, the ventral striatum (O'Doherty et al., 2003), the lateral habenula (Salas & Montague, 2010) or the midbrain (D'Ardenne, McClure, Nystrom, & Cohen, 2008). This was true at the whole brain and ROI level. Either these areas do not respond at all to option level PEs, or their response is so small that cannot be observed. However, Diuk et al. (2013) do find responses in the ventral striatum integrating information across an option, whereas no responses in dACC. The study involved extended PEs, computed at the end of the option, spanning information about the accrued rewards during the option. Even though these were RPEs, it means that VS *is* receiving information at the option-level and is not solely responding to changes in flat RPEs. One possible nullifying factor is the spectrum of preferences in the population. As illustrated in Figure 3.7, if there is a null-centered spectrum of true preferences, an area with an unsigned response will reflect both the aversive and appetitive nature of a jump, in statistical analysis at the population level. However, an area with a signed response will mirror the distribution of preferences around 0. Unlike the behavioral studies, our fMRI analysis rely strongly on population level tests, and could make responses in VS undetectable.

Given that the PPE is assumed to arise from hierarchical processing, it may appear necessary for us to have established independent of the imaging experiments that subjects represent the delivery task hierarchically. We have claimed that the imaging data provide evidence both for the PPE and for the logically prior proposition that the delivery task is performed hierarchically. Isn't there necessarily some circularity in this analysis? Despite the appeal of this intuition, there is in fact nothing circular in our interpretation of the data. To show this formally, let us define the following terms:

A: The event that the task is represented hierarchically

$\overline{A}$: The event that the task is not represented hierarchically

B: The event that the task gives rise to a PPE

$\overline{B}$: The event that the task does not give rise to a PPE

D: Our neuroimaging findings.

On purely logical grounds, it is clear that:

$B \rightarrow A$: If B were true, then A would necessarily also be true

$\overline{A} \rightarrow \overline{B}$: If A were false, then B would necessarily also be false

Given these two premises, basic probability yields the following two conclusions:

$$P(B \mid D) = \frac{P(D \mid B)P(B \mid A)P(A)}{P(D \mid B)P(B \mid A)P(A) + P(D \mid \overline{B})P(\overline{B})} \tag{3.3}$$

$$P(A \mid D) = \frac{P(D \mid B)P(B \mid A)P(A) + P(D \mid A \cap \overline{B})P(A \cap \overline{B})}{P(D \mid B)P(B \mid A)P(A) + P(D \mid \overline{B})P(\overline{B})} \tag{3.4}$$

Equation 3.3 gives the posterior probability of the PPE hypothesis, given the neuroimaging data. Equation 3.4 gives the probability of hierarchical processing, given those same data. Two points are worth noting. First, there is no circular or reciprocal dependency between the two equations.[4] Given the appropriate likelihoods and prior probabilities, the equations can be evaluated in parallel. It is thus logically sound to draw parallel conclusions from the imaging data concerning both hierarchical processing and the

[4]The two expressions do of course share terms, and will thus be correlated, but this is no indication of circularity or tautology As an aside, also note that $P(B \mid D) = P(A \cap B \mid D)$; our experiment may be seen as evaluating the *joint* hypothesis $A \cap B$.

PPE. Second, both probabilities depend inversely on $P(D \mid \overline{B})$, the probability that the data might have been obtained in the absence of a PPE. This indicates the importance of ruling out alternative explanations for the imaging results. It is here that the behavioral study comes in, since it rules out an interpretation of the imaging data based on primary reward at subgoal. Naturally, both probabilities, $P(B \mid D)$ (Equation 3.3) and $P(A \mid D)$ (Equation 3.4), also depend on $P(A)$, the a priori probability that the delivery task is performed hierarchically. Previous research makes it reasonable to consider this probability to be fairly high: As we have recently reviewed elsewhere (Botvinick, 2008; Botvinick, Niv, & Barto, 2009), decades of research in cognitive psychology (e.g., Miller et al., 1960; Cooper & Shallice, 2000; Zacks, Speer, Swallow, Braver, & Reynolds, 2007), developmental psychology (e.g., Saffran & Wilson, 2003), neuropsychology (e.g., Schwartz et al., 1995; Badre et al., 2009), functional neuroimaging (e.g., Koechlin et al., 2003; Badre & D'Esposito, 2007), and neurophysiology (e.g., Fuster, 2001) indicate that hierarchical representation is ubiquitous, and perhaps even obligatory in human behavior. The possibility that our experimental task, with its very salient goal-subgoal structure, might constitute an exception to this general rule seems improbable. Nevertheless, the importance of the hierarchy assumption prompted us to consider whether our data might provide some additional, independent and convergent evidence for hierarchical processing.

One opportunity, in this regard, is suggested by recent neurophysiological research, which has discovered phasic activity within the dorsolateral prefrontal cortex and dorsolateral striatum at sequence boundaries (Barnes et al., 2011; Fujii & Graybiel, 2003; Jin & Costa, 2010). We reasoned that, if participants in our experiment represented the delivery task hierarchically, such activity should occur at the point of subgoal attainment, since this marks the completion of one subsequence and the onset of another. Importantly, the moment of subgoal attainment in our task also requires a

shift in visual attention; to control for this factor, we used package-jump events (pooling across jump types E and D) as a baseline, since these events also require a shift in visual attention but do not lie at a subtask boundary. The resulting contrast revealed relative activation at subgoal attainment ($p < .01$, corrected as in previous analyses) at three points within dorsolateral prefrontal cortex (Talairach coordinates: 63, 7, 25; -61, 4, 30; and -51, 40, 19) and bilaterally within dorsolateral striatum (15, -14, 25; -12, 11, 19). Relative activation was also observed in left anterior parietal cortex spanning the intraparietal sulcus, in the right precuneus, in bilateral middle occipital gyri, and in the cerebellum. Interestingly, the prefrontal areas identified in this contrast lie near to areas identified in recent neuroimaging studies aimed at isolating regions responsible for instantiating hierarchical representations of action (Koechlin et al., 2003; Badre & D'Esposito, 2007). We refrain from drawing strong conclusions from this apparent correspondence, given the many differences between the task and analysis employed here and ones involved in those previous studies. However, the finding of phasic activation in these frontal regions at the subtask boundary within our task does appear to offer some convergent support for our assumption that participants represented the delivery task in a hierarchical fashion.

Overall, these findings are consistent with the dACC having a role in learning through a Pearce-Hall model, at the level of subgoals. What does this mean? In the study where people could learn the association between button presses and direction of jumps of the subgoal, while preserving goal distance, and carry on this choice, we could expect dACC, but not VS, or lateral habenula, response during learning. As learning happens, and the association between presses and subgoal locations becomes predictable, responses in the dACC should be reduced and eventually inexistant.

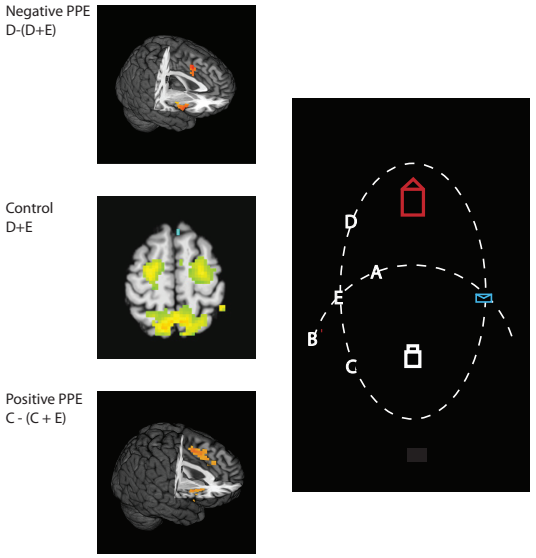

Figure 3.6. Whole-brain results for negative, and positive PPEs. (Negative PPE) Contrast of jumps type D - $(D + E)$. Shown are regions displaying a positive correlation with the PPE, independent of subgoal displacement. Talairach coordinates of peak are (0, 9, 39) for dACC, and (45, 12, 0) for right anterior insula. Not shown are foci in left anterior insula (-45, 9, -3) and lingual gyrus (0, -66, 0). (Control) Axial view (z = 53) of the BOLD activity for events D and E ($p < .01$ corrected) contrasted with no jump condition. Talairach coordinates for the peak voxel for the clusters shown are (18, -66, 51) intraparietal sulcus, (-27, -12, 54) and (24, -15, 51) for frontal eye fields. (Positive PPE) Contrast of second fMRI experiment, using type C - $(C + E)$. Shown are regions displaying a positive correlation with the PPE, independent of subgoal displacement, which overlapped with regions for Negative PPE. $p < .01$, corrected using cluster size. Color indicates general linear model parameter estimates, ranging from 3.0 x 10-4 (palest yellow) to 1.2 x 10-3 (darkest orange).

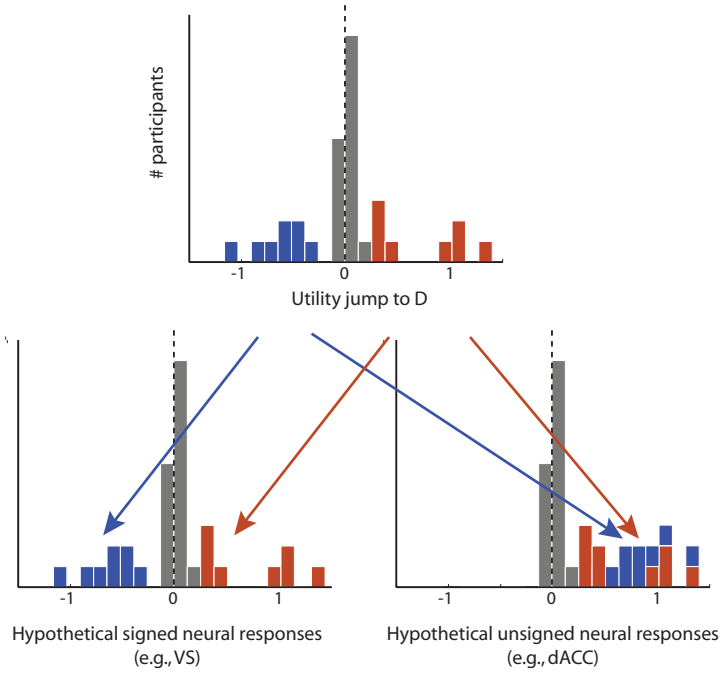

Figure 3.7. Effect of a spectrum of preferences around 0 on the detection of neural PE responses. Assuming the population of participants had a similar distribution of preferences as the one observed in the last behavioral experiment, this will undermine the detection of a signed response (left). However, in an unsigned case, both extremes of preferences contribute to an increase in neural activity (right).

Chapter 4

Neural Correlates of Pseudo-Reward and Reward Prediction Errors

4.1 Chapter Summary

In this last experimental chapter, we describe one fMRI study which aimed at comparing neural responses to PPEs and RPEs. In addition to PEs related to subgoal jumps, there were monetary RPEs at the end of each trial. These were introduced to further ground the comparison between RPEs and PPEs.

- Participants played a similar spatial delivery paradigm. Two-thirds of the trials involved a jump of the subgoal. All jump trials elicited both an RPE and a PPE. The spatial distribution of jumps was specially designed to uncorrelate RPEs and PPEs, as well as the distance between the old and new subgoal location.

- We observed an unsigned dACC response to RPEs, elicited by subgoal jumps. This is consistent with research showing absolute responses in this area.

- However, in contrast with the findings from the first neuroimaging studies, there was no cingulate response to absolute PPEs. Though surprising, this is consistent with the mutually exclusive pattern of choices we observed in the behavioral studies: participants' choices only reflect subgoal distance *when* there is no change in overall distance.

- We replicated VS responses to positive RPEs, driven by unexpected monetary outcomes. There was no response for negative RPEs. No striatal response was elicited in PEs related to subgoal jumps. In contrast with research showing VS-dACC co-activation for RPEs, no dACC activity was observed for monetary outcomes.

- The dissociation between VS and dACC points to the possibility that PEs related to subgoal manipulations in our task, and dACC activity

in this thesis, are actually related to violations of transitions (state prediction errors), but not of reward predictions. Thorough analyses excluded that dACC responses would be due to spatial shifts of attention.

- Overall, we observed evidence for a process of hierarchical prediction in dACC. It is a matter for future research whether these responses drive updating of transitions (as predicted by this last set of findings), or of hierarchical values (as dictated by our initial predictions). In the general discussion, we present an experiment seeking to disambiguate between these two possibilities.

4.2 Introduction

According to our initial hypothesis, the structures that respond to RPEs also encode RPEs. This was based on a parsimonious extension of RL to HRL, and on recent study (Diuk et al., 2013), which found ventral striatal responses to prediction errors at two levels of hierarchy. In order to directly address this hypothesis, we tested RPEs and PPEs using the same hierarchical spatial paradigm. The predictions were that 1. RPEs can be elicited using the delivery task — given the clear pattern of choices shown in the behavioral chapter, 2. PPEs arise in the same region, and 3. unsigned PEs should be observed in dACC. In addition, 4. we benchmark jump RPEs against probabilistic monetary rewards, as these have a strong prior for robust responses in regions involved in RPEs (Niv, 2009).

4.3 An fMRI Experiment Crossing Valence and Level of Hierarchy

Methods

As a recapitulation of our paradigm, we can elicit different types of PEs by having the subgoal unexpectedly jump to different points in space. As shown in Figure 4.1, jumps on the ellipse preserve overall distance and only change action costs to the subgoal (C - decrease in distance, positive PPE, D - increase in distance, negative PPE). Jumps to points A and B change overall distance, but not initial distance, and thus only trigger positive and negative RPEs, respectively. Because a paradigm with five jump conditions, including a jump to point E, and a non-jump condition would be infeasible either in terms of power or duration, we set for a paradigm where all jumps involved a PPE and an RPE, which were parametrically, but not categorically, uncorrelated.

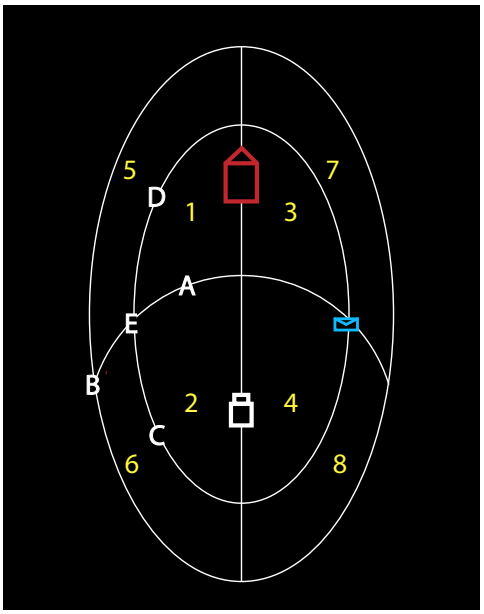

Figure 4.1. Types of PEs. The previous paradigms used jumps to locations D and E, or C and E. In this experiment PEs were elicited by having the subgoal jump to random points in eight regions of space, highlighted in yellow: 1-4 - positive RPEs, 5-8 - negative RPEs, odd - negative PPEs, even - positive PPEs. Locations A and B depict RPEs without PPEs, C and D PPEs without RPEs, and location E should not trigger any PE (except for salience).

Participants. Forty participants were recruited from the Princeton University community (range 18-27 years, $M = 20$, $SD = 1.78$, 15 male, 38 were right-handed and 2 were left-handed, joystick was always held in the right hand). 8 participants were excluded, totalling 48 recruited participants (7 for head movement larger than 2.5 *mm* and 1 for failure to complete the task on time). All participants received monetary compensation at a departmental standard rate, and a monetary bonus for performance plus a probabilistic payment described as a tip, as detailed below.

Materials, task and procedure. The task consisted of three parts: a short behavioral pratice outside the scanner, for 12 trials, using a joystick held in the right hand (Logitech International, Romanel-sur-Morges, Switzerland) preceded a practice in the scanner, a 12 trial practice inside the scanner, using an MR compatible joystick (MagConcept, Redwood City, CA) during structural scan acquisition and a third phase of 132 trials (6 runs of 22 trials) for approximately sixty minutes, where functional data were collected. At the beginning and end of each run a central fixation cross was displayed for 10000 ms. The average run length was 11.73 minutes.

Participants played a variant of the delivery task. On each trial truck, envelope and house occupied the vertices of a virtual triangle with vertices at pixel coordinates (-90, 320; truck), (150, 0; envelope) and (0, -200; house) relative to the center of the screen (resolution 1024 x 768 pixels), but assuming a random new rotation at the onset of each trial. The task was to move the truck first to the package and then to the house. Each joystick movement displaced the truck a fixed distance of 50 pixels. The initial location of the truck was determined such that it would be at 3 optimal steps of distance (50 pixels) from the planned location for jumps to happen (0,200). At this location the envelope was equidistant from the truck and goal, to allow for equal variance in both positive and negative prediction errors.

Because of variance in performance, participants would never fall exactly on the planned point (0,200). The jump happened when the truck was closer than 250 pixels to the envelope or 400 pixels to the house — this approximated a line. When the truck passed these boundaries, a brief tone was played, the truck and envelope would flash yellow, and joystick movements were ignored for 900 ms. In one-third of the trials the envelope would stay in the same location. In the remaining two-thirds it would jump to a new location (see the next paragraph for details on the jump locations). Participants were told that the envelope sometimes stayed in the same place, and sometimes it jumped. No information was given about

the location of the jump. We emphasized that there was no contingency between performance and the probability of jumping.

Post-jump envelope locations were determined *a priori* using a Monte Carlo approach. The space of (x,y) coordinates was sampled to yield an equal number of positive and negative RPEs and PPEs, and ipsilateral and contralateral jumps. In addition, we bounded negative RPEs. Negative RPEs are only restricted by the screen boundaries, whereas the maximal positive RPE is a jump to the straight line between truck and house. Datasets were constrained to have a maximal negative RPE of the same magnitude as the maximal PPE. Figure 4.1 illustrates each of these areas of space in an example dataset. After sampling within these boundaries, we selected datasets that (1) had a mean PPE approaching zero (mean of PPE distance less than half a standard deviation away from the mean of the set of samples, which was zero), (2) had a mean RPE approaching zero (mean of RPE distance less than a third of a standard deviation away from the mean of the set of samples, which was zero), (3) had a low sum of absolute correlation between variables (was farther than minus one standard deviation away from the mean of the sum of the pairwise correlations between PPE, RPE and jump distance), and (4) had a high variance (datasets with a standard deviation more than one standard deviation away from the mean of standard deviations; this counteracted the bias for low variability from the previous conditions). Out of the remaining datasets we randomly sampled one. Within this dataset we used the same Monte Carlo approach to look for possible orderings of trials that could allow for an exploration of values. However, PEs from a model with learning ($\alpha = .1$) were highly correlated with those from a model which only reflected the current trial ($\alpha = 1$). These sampling and selection procedures were repeated for each participant and for each task phase.

As mentioned before, because of errors in performance, the jump was triggered at a truck location that approximated, but not equaled, the

planned truck location. To ensure that performance would not grossly change the correlations and the means of PEs significantly, we tested the selected 40 datasets with an artificial agent with the same accuracy that was observed in the previous behavioral tasks. Indeed, for the actual datasets, taking into account participants' performance, across jump conditions, the average RPE was close to zero, $M = .1$ steps, converting distance in pixels to steps, though with a relatively large variability, mean $SD = 1.66$ steps, and a mean maximum of 3.19 steps; and the same for the average PPE ($M = 0$, mean $SD = 2.02$ steps, mean maximum = 4.44 steps). It should be noted that the means for individual runs could be different from zero. This was to discourage participants from tallying how many types of each event had happened in a run. Changes in local distance, between pre and post-jump, were on average 4.41 steps ($SD = 1.7$). The correlation between PPE and RPE was .31, correlation between PPE and jump distance 0, correlation between RPE and jump distance -.37.

After the jump, participants headed towards the new location of the subgoal. When the truck passed within 30 pixels of the package, the package moved to the truck and remained there for the subsequent moves. When the truck with the package passed within 30 pixels of the house, the truck with the package appeared within the house. This image was displayed for 200 ms. After this period, a screen was shown with monetary information, as shown in Figure 4.2.

Participants were paid a flat rate of 150 delivery bucks, a task currency that would be converted to dollars. Though they were not told what the conversion rate was, they were told that if they "worked hard a maximum of $12" could be attained at the end of experiment in addition to the departmental rate. In order to emphasize the cost of distance, gas was deducted from the flat rate. This was .1 delivery bucks per *actual* pixel travelled (truck at the start - truck at the jump - truck at package pick up - truck in the house), up to a maximum of 100 delivery bucks. This was accompa-

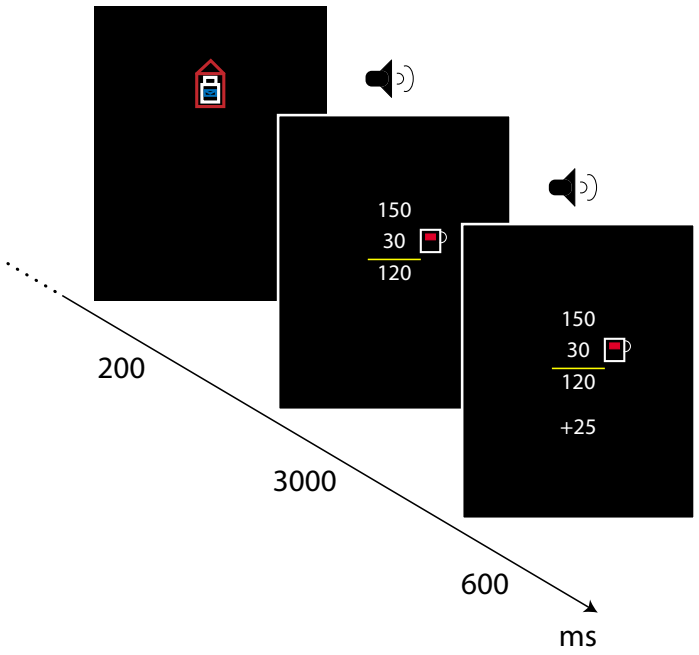

Figure 4.2. Eliciting RPEs through monetary outcomes. After delivering the envelope to the house, the truck would be shown inside the house for 200ms. After this period, a brief tone was played and a breakdown of payment and distance costs was shown for 3000ms, followed by a screen with probabilistic monetary payment (+25, 0 or -25), accompanied by a tone consistent with the valence (coin sound, neutral tone, or sad trumpet).

nied by the sound of cash register. After 3000 ms, a probabilistic monetary reward appeared at the bottom of the screen (see Figure 4.2). This was introduced to compare reward prediction errors arising from package jump with reward prediction errors from monetary reward. Participants could get 25, -25 or 0 delivery bucks with equal probability. They were told that this was not contingent on their performance but that it was worthwhile to pay attention to this additional payment, given that final payment was a sum of rewards accrued during all task phases. To ensure attentional capture,

we introduced a sound at the moment of this information (coin sound for 25, different from the one for the flat rate, a sad trumpet sound for -25, and a brief tone for 0, all sounds had the same 100 ms duration). This was displayed for 600 ms and was followed by a fixation cross that remained on screen for 700 ms. At the end of each run participants would be given a self-paced break.

Image acquisition. Data were acquired with a 3T Siemens Skyra (Malvern, PA) MRI scanner using a sixteen-channel head coil. High-resolution (1 mm^3 voxels) T1-weighted structural images were acquired with an MP-RAGE pulse sequence at the beginning of the scanning session. Functional data were acquired using a high-resolution echo-planar imaging pulse sequence (3 x 3 x 3 mm voxels, 35 contiguous slices, 3 mm thick, interleaved acquisition, TR of 2000 ms, TE of 30 ms, flip angle 90 °, field of view 192 mm, aligned with the Anterior Commissure - Posterior Commissure plane). The first five volumes of each run were ignored.

Data analysis. Data were analyzed using AFNI software (Cox, 1996). The T1-weighted anatomical images were aligned to the functional data. Functional data was corrected for interleaved acquisition using Fourier interpolation. Head motion parameters were estimated and corrected allowing six-parameter rigid body transformations, referenced to the initial image of the first functional run. A whole-brain mask for each participant was created using the union of a mask for the first and last functional images. Spikes in the data were removed and replaced with an interpolated data point. Data was spatially smoothed until spatial autocorrelation was approximated by a 6 mm FHWM Gaussian kernel. Each voxels signal was converted to percent change by normalizing it based on intensity. The mean image for each volume was calculated and used later as baseline regressor in the general linear model, except in the region of interest analysis where the mean image of the whole brain was not subtracted from the data. Anatomical images were used to estimate normalization parameters to a template in Talairach

space (Talairach & Tournoux, 1988). These transformations were applied to parameter estimates from the general linear model.

General linear model analysis. For each participant we created a design matrix modeling experimental events and including events of no interest. At the time of an experimental event we defined an impulse and convolved it with a hemodynamic response. The following regressors were included in the model: (a) an indicator variable marking the occurrence of all auditory tone / package flash events, (b) an indicator variable marking the occurrence of all jump events, (c) a parametric regressor indicating the change in distance to subgoal induced by each jump, mean-centered, (d) a parametric regressor indicating the change in distance to goal induced by each jump, mean-centered, (e and f) indicator variables marking subgoal and goal attainment, (g) an indicator variable marking all periods of task performance, from the initial presentation of the icons to the end of the trial, (h) an indicator variable for delivery of monetary reward (encompassing the positive, 25, negative, -25, and neutral, 0, events), (i) an indicator variable for the positive reward, 25, and (j) an indicator variable for the negative reward, -25. Also included were head motion parameters, and first to third order polynomial regressors to regress out scanner drift effects. A global signal regressor was also included. In additional analyses, instead of indicator variables encompassing signed positive and negative events, we separated regressors for positive negative events, or included them in a unsigned way, with one regressor for the jump PEs and one regressor for the monetary PEs. All parametric regressors were mean-centered after all changes.

Group analysis. For each regressor and for each voxel we tested the sample of 40 subject-specific coefficients against zero in a two-tailed t test. We defined a threshold of $p = .01$ and applied correction for multiple comparison based on cluster size, using Monte Carlo simulations as implemented in AFNIs AlphaSim. We report results at a corrected $p < .01$.

Region of interest analysis. For the first fMRI experiment we defined ventral striatum (including the olfactory tubercle) based on anatomical boundaries on a high-resolution T1-weighted image for each participant. Mean coefficients were extracted from this region for each participant. Reported coefficients for all regions of interest are from general linear model analyses without subtraction of global signal. The sample of 40 subject-specific coefficients were tested against zero in a two-tailed t test, with a threshold of $p < .05$.

Results

Behavior

Participants completed a trial on average within 19.81 steps ($SD = 4.31$). The average step for the pause was 5.57 ($SD = .49$). After a jump, participants reacted within 1.48 s ($SD = .18$) and accuracy was 74.5 ° ($SD = 8.02$). As expected, responses in the jump condition were significantly slower and less accurate than in the no jump condition (RT: mean difference = 67 ms, $p = .002$; Accuracy = .08 °, $p < .001$). This effect was larger in jumps that involved a larger distance between the two subgoals location ($\rho = .09$ for RTs, $p < .001$, and $\rho = .07$ for accuracy, $p < .001$). There was no significant effect on RT or accuracy of RPE ($p = .76$ and $p = .16$ respectively). There was a small positive trending relationship between magnitude of PPE and accuracy ($\rho = .04$, $p = .07$), and no significant correlation with RT ($p = .45$).

fMRI

Jump-related PEs. Across all jumps, we found a robust parametric effect of jump distance in bilateral frontal eye fields and in posterior parietal cortices. Unsigned RPEs yielded dorsal anterior cingulate activity in a region

similar to the unsigned PPE in the previous studies, Figure 4.3. In contrast, we observed no medial response for unsigned PPEs. There were responses in areas for which we had no *a priori* hypotheses for: increase in BOLD in middle frontal gyrus (BA 8), and bilateral medial temporal gyrus (see Table 4.1 for coordinates and cluster sizes). A separation of RPEs into positive and negative yielded no significant response at the whole-brain level. We observed activity for signed PPEs several temporal and occipital areas (Table 4.1).

Replicating the findings from our previous studies, the jump manipulation (independent of RPE or PPE), elicited robust responses in FEF, posterior parietal cortices, areas related to spatial shifts of attention (Corbetta et al., 2008), as well as a decrease in ventromedial PFC, posterior cingulate and retrosplenial cortex (Raichle et al., 2001) — areas whose joint activity, as the default mode network, often fluctuates inversely with shifts of attention and task engagement.

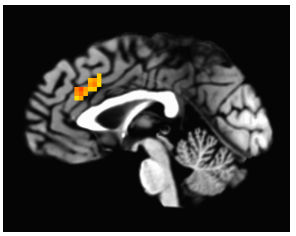

Figure 4.3. Effect of unsigned RPE on medial prefrontal cortex. Peak coordinates are $(1.5, 28.5, 29.5)$, $p < .05$ corrected.

Monetary PEs. Positive RPEs yielded a trending increase in BOLD signal in the right ventral striatum (-16,2,-4; $p = .08$ at whole-brain level) — the independent ROI analysis yielded the same result of a significant response in right striatum, $p < .05$. Responses in the left striatum were not significant ($p = .22$). No cingulate response was observed for positive or negative RPEs at a whole-brain level. There were large differences in transverse temporal cortex, across positive and negative RPEs, which likely

reflected differences in the auditory stimuli, given the difference in intensity for the loss and gain sound and the location in auditory cortex (see Table 4.2 for cluster details).

Table 4.1. Whole-brain clusters for jump-related regressors ($p < .05$, corrected by volume; size in voxels; t value and coordinates for peak voxel; R. = Right, L. = Left, S. = Superior, g. = gyrus)

Regressor/Area	**BA**	**size**	***t***	**x, y, z**
Absolute RPE				
L. dACC	32	48	3.98	1.5, -28.5, 29.5
Positive RPE				
-	-	-	-	-
Negative RPE				
-	-	-	-	-
Absolute PPE				
R. middle temporal g.	19	214	4.6	-34.5, 76.5, 20.5
R. fusiform g.	17	59	4.4	-25.5, 50, -6.5
Positive PPE				
R. middle temporal g.	19	372	4.99	40.5, 79.5, 20.5
Negative PPE				
R. precuneus	7	84	-4.62	-10.5, 58.5, 35.5
R. middle occipital g.	18	53	4.14	-40.5, 79.5, -6.5
R. middle temporal g.	22	48	-4.6	-50, 10.5, -6.5
Jump				
R. precuneus	7	4452	13.5	-4.5, 67.5, 47.5
S. temporal g.	41	938	-6.5	46.5, 31.5, 17.5
R. middle frontal g.	6	756	8.90	-25.5, 7.5, 53.5
R. medial frontal g.	8	745	-4.92	-7.5, -40.5, 38.5
L. postcentral g.	3	395	-6.16	37.5, 31.5, 60
R. middle frontal g.	9	221	7.32	-28.5, -31.5, 30

continued on next page

continued from previous page

Regressor/Area	**BA**	**size**	***t***	**x, y, z**
L. middle frontal g.	47	198	-4.67	34.5 , -34.5, -3.5
L. lentiform nucleus	-	189	-4.85	25.5, 4.5, -6.5
R. postcentral g.	40	162	-5.40	-58.5, 20, 17.5
R. superior temporal g.	22	141	-4.87	-55.5, 7.5 , 8.5
L. middle frontal g.	8	138	6.09	34.5 , -28.5, 38.5
L. dorsal caudate	-	118	6.78	16.5 , -7.5, 14.5
R. culmen	-	116	-4.5	-1.5 , 55.5 , -15.5
L. parahippocampal g.	37	98	5.72	28.5 , 46.5 , -9.5
L. cingulate g.	24	89	-4.98	4.5 , 13.5 , 35.5
L. posterior cingulate	23	84	-4.04	4.5 , 49.5 , 23.5
R. lentiform nucleus	-	70	-4.21	-20 , 4.5 , -6.5
R. middle frontal g.	47	68	-4.50	-34.5, -31.5 , -3.5
R. lentiform nucleus	-	57	4.92	-19.5, -13.5, 5.5
R. parahippocampal g.	37	54	4.62	-28.5 , 43.5 , -6.5
R. cuneus	18	48	-3.64	-4.5 , 76.5 , 17.5
Jump distance				
R. precuneus	31	86	4.3	-7.5 , 43.5, 44.5
R. middle frontal gyrus	6	66	4.3	-19.5, -4.5, 59.5

Table 4.2. Whole-brain clusters for monetary outcomes regressors ($p < .05$, corrected by volume; size in voxels; t value and coordinates for peak voxel; R. = Right, L. = Left, S. = Superior, g. = gyrus)

Regressor/Area	**BA**	**size**	***t***	**x, y, z**
Absolute RPE				
R. posterior insula	13	149	5.02	-43.5 , 13.5 , 3.5
L. posterior insula	13	108	6.95	43.5 , 19.5, 0
Middle occipital g.	17	96	-4.53	31.5 , 61.5 , 0

continued on next page

continued from previous page

Regressor/Area	**BA**	**size**	***t***	**x, y, z**
R. middle temporal g.	21	59	4.63	-50, -4.5, -10
R. fusiform g.	37	56	4.43	-34.5 , 43.5 , -10
R. cuneus	18	52	-4.35	-16.5 , 76.5 , 26.5
R. inferior frontal g.	9	49	4.25	-34.5 , -4.5 , 30
Positive RPE				
Transverse temporal g.	41	362	-6.97	46.5 , 20 , 11.5
Transverse temporal g.	41	350	-5.65	-46.5 , 20 , 11.5
R. fusiform g.	37	203	5.7	-34.5, 43.5, -10
L. fusiform g.	37	92	5.6	28.5, 49.5, -10
Negative RPE				
R. superior temporal g.	22	1324	8.67	-46.5 , 13.5 , 0
L. superior temporal g.	22	788	9.06	46.5 , 16.5 , 8.5
R. cuneus	18	249	-5.14	-16.5 , 76.5 , 20.5
L. middle occipital g.	19	210	-5.27	31.5 , 61.5, 0
L. Cuneus	17	188	-5.72	16.5 , 85.5 , 11.5
R. inferior temporal g.	37	96	-4.55	-46.5, 61.5, -0.5
L. thalamus	-	85	4.75	4.5, 16.5 , -3.5
L. dorsal caudate	-	75	-5.25	7.5 , -13.5 , 11.5

4.4 Chapter Discussion

The aims of this experiment were to compare neural responses to RPEs and PPEs within subjects and using the same spatial paradigm. PEs were elicited by unexpected changes of subgoal location, while executing In addition, we included RPEs triggered by monetary outcomes to further the comparison with PPEs.

We observed an unsigned reward prediction error in dorsal anterior cingulate cortex (dACC), an increase in BOLD activity with the magnitude,

but not the valence, of the RPEs. Unsigned prediction errors are part of learning models driven by surprise (Pearce-Hall Mackintosh, 1975; Pearce & Hall, 1980; Pearce et al., 1982). This is consistent with previous studies using single-unit recordings(Hayden et al., 2011), fMRI (Jessup, Busemeyer, & Brown, 2010), or EEG (Talmi, Atkinson, & El-Deredy, 2013), which also found unsigned responses in dACC, as recently reviewed in Roesch et al. (2012). ACC could be responsible for driving learning by association (together with amygdala, as proposed in Roesch et al., 2012), or use this deviation for re-evaluation of control (Shenhav, Botvinick, & Cohen, 2013), or option policies (Holroyd & Yeung, 2012).

Alternative interpretations of this response are conflict between courses of action, and errors (Botvinick, Braver, Barch, Carter, & Cohen, 2001; Holroyd et al., 2004; Ridderinkhof, Ullsperger, Crone, & Nieuwenhuis, 2004, for a review). However, RPEs and PPEs were not significantly correlated with either reaction times or accuracy, making it unlikely that the response we observed was due to these two alternative factors.

Surprisingly we did not replicate the cingulate response to PPEs. What might be driving this mutually exclusive response to PPEs or RPEs? Previous studies have shown that people are able to simultaneously process reward information at different levels of abstraction (Krigolson & Holroyd, 2007; Badre et al., 2010; Diuk et al., 2013). Therefore, this suggests that the source of competition in our task is not at the level of decisions. It is telling that the spatial nature of our paradigm is the most evident difference with previous hierarchical paradigms namely from that of Diuk et al.. On this note, research on global *vs.* local perceptual processing (Navon, 1977), and spatial frames of reference (Behrmann & Tipper, 1999), comes to bearing. Navon has shown that, in stimuli with both global and local features (e.g., a large H composed of smaller s), participants exhibit interference from global information (identity of large letter) when responding at a local level (identity of smaller letters), but not the reverse. In Behrmann and Tip-

per (1999) it is shown that certain types of spatial information cannot be processed simultaneously. The hypothesis is thus that participants cannot judge both overall and local distances simultaneously, and that reward incentivizes processing of overall distance. This would be consistent with our behavioral studies, whereby people were only sensitive to subgoal-related action costs when no evaluation of overall action costs was necessary. In the general discussion, we present a paradigm to test this spatial hypothesis.

It was also surprising to observe a dissociation between ACC and VS in jump RPEs. Both receive prominent dopaminergic input from the midbrain (Szabo, 1979; Miller & Vogt, 2009), and are heavily interconnected (Berendse, Graaf, & Groenewegen, 1992; Parkinson, Willoughby, Robbins, & Everitt, 2000; Croxson et al., 2005; Krebs, Boehler, Roberts, Song, & Woldorff, 2012). Moreover, these two regions have been extensively reported to be co-activated in studies examining neural correlates of RL (Croxson, Walton, O'Reilly, Behrens, & Rushworth, 2009; Walton et al., 2009; Botvinick, Huffstetler, & Mcguire, 2009; Krebs et al., 2012), and particularly RPEs, in rodent, and primate research (O'Doherty et al., 2003; Ullsperger & von Cramon, 2003; Walton, Devlin, & Rushworth, 2004; Amiez, Joseph, & Procyk, 2005; Mars et al., 2005; Haruno & Kawato, 2006; Seo & Lee, 2007; Rutledge, Dean, Caplin, & Glimcher, 2010; Hayden et al., 2011), though not with exceptions (Holroyd et al., 2004; Viard, Doeller, Hartley, Bird, & Burgess, 2011; Diuk et al., 2013, in this last study, ACC, among other areas, was observed to be active in spatial violations, though the reward structure was not as clear as in our study). In spite of the clear behavioral preferences for closer goals, it is possible that the observed response is not directly reward-related, but to violations of outcomes, in other words a state prediction error. Detection of these violations is actually a core component of some theories of ACC function (PRO theory, Alexander & Brown, 2011). This would be consistent with two recent integrative theories of ACC, which stipulate that any signal that requires

re-evaluation of the amount of cognitive control (Shenhav et al., 2013) or of temporally-extended actions (Holroyd & Yeung, 2012). In the next chapter, as part of future directions, we propose a learning paradigm which seeks to elucidate the nature of the observed medial frontal response.

Chapter 5

General Discussion

5.1 Overview of Empirical Findings

The aim of this thesis was to explore behavioral and neural correlates of hierarchical reinforcement learning (HRL), not explainable by a flat RL model. We used a spatial paradigm that was divisible into two subtasks, each composed of a sequence of actions. From the behavioral studies we observed several properties of a hierarchical agent:

- Values exist at multiple levels of a task — preferences were revealed at the root and option level.
- Values at the root level dominate values at option level — in the presence of a trade off between reward and pseudo-reward, participants overwhelmingly chose to maximize reward.
- Option-level values were expressed during option execution but not while executing a different option or a root-level policy.

At a neural level, we sought to test whether the same structures involved in coding root-level prediction errors (RPEs) would respond to option-level prediction errors (PPEs). In particular, our main prediction was for a consistent engagement of midbrain dopamine afferents. We found confirmations of an HRL process at play, though equivocal evidence for involvement of the structures that code for RPEs:

- Dorsal anterior cingulate cortex (dACC) responds to option-level prediction errors (PPEs) in a fashion similar to RPEs, evidenced by metabolic and electrophysiological correlates.
- dACC responds to prediction errors in an unsigned way, consistent with an involvement in learning driven by surprise (Pearce-Hall model).

- No striatal response was observed to PPEs. In contrast, there was a response to monetary RPEs. No consistent habenular or amygdala response was detected.

- Evidence is not suggestive of a role of dopamine in coding PPEs.

In the next sections we discuss specific and general future directions of the work presented in this thesis. In addition, we compare the theoretical scaffold of this thesis, with other theoretical proposals and corresponding empirical findings. Finally, we address issues pertinent to an implementation of HRL that have not been the focus of the thesis: the problem of subgoal discovery, and model-based *options*.

5.2 Future Directions

Specific directions

This thesis leaves several issues open for further research. Broadly they concern explaining dissociations we observed, thorough explorations of dopaminergic function in hierarchical domains, and assessment of further HRL predictions.

Spatial determinants of attention at several levels of hierarchy. We observed that participants either chose taking into account goal distance or subgoal distance, but not both. Our main hypothesis is that such dissociation was due to incompatibility of processing global *vs.* local information. In order to ascertain such hypothesis, it would be informative to design a psychophysical study, with no task structure or reward. If competition were observed in a purely perceptual paradigm, it would give support to the idea that the absence of neural PPEs in the last fMRI study was due to impaired spatial processing.

The proposed experiment uses the same spatial locations as in the last fMRI study (see Figure 5.1). However, there is no cover task: participants

see three isoluminant icons on a screen, which are only different in their shape and color. In two-thirds of the trials the middle vertex (blue circle) jumps to a new location (see Figure 5.1A and B). For 900 ms participants see the old and new location of the vertex. After this period, it disappears and participants have to estimate the extent to which the either the overall or the "subgoal" distance changed (see Figure 5.1C). In the remaining third of the trials, there will be no jump though participants still have to estimate the amount of change.

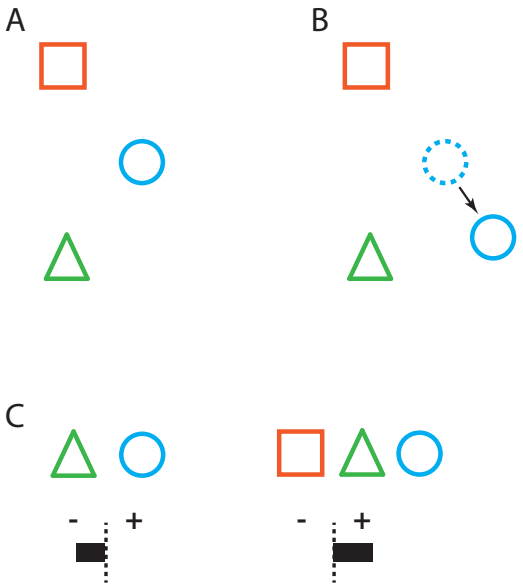

Figure 5.1. Estimation of global and local distances. Participants play a distance estimation task between two changing displays. All locations will be the same as in the last fMRI study. (A) Pre-jump display. (B) Jump of the middle vertex to a new location (equivalent to a jump of a subgoal). (C) Estimation of change of local (left) or overall (right) distances using a slider bar.

We predict that under situations where overall distance changes, participants will be insensitive to changes in local distance. As an additional manipulation, we can reward each distance discrimination differentially and assess how does that affect estimation of the other distance.

PPEs in a learning situation. The PPEs in the presented studies were elicited in a Pavlovian-like probabilistic setting. However, PPEs should be correlated with future approach behavior as dictated by TD.

We propose a study where both levels of the task hierarchy drift independently. Initially, participants become acquainted with the delivery task, with a single subgoal. In the test phase, they can choose between 3 different subgoals, illustrated in Figure 5.2A. They do so by entering the respective colored area on their first step. After the first step the unchosen subgoals disappear. Each of the subgoals is characterized by independent, random, trajectories of PPEs and RPEs through time, as shown in Figure 5.2B. Trials with forced exploration will be introduced.

This experiment would allow to disambiguate the nature of dACC response we observed in the three neuroimaging studies. In addition, it would provide a better exploration of PPEs, as we can obtain the fits for models with or without V_o based on the choice behavior (similarly to Diuk et al., 2013).

Manipulating task structure independently of reward-based computations. The purpose of this experiment is to independently manipulate task structure, and observe PPEs based on the putative structure. The independent manipulation of subtasks will be achieved through prior exposure to statistical structure, as shown in Figure 5.3A, Exposure Phase. There is ample evidence that people are sensitive to such structure (Turk-Browne & Scholl, 2010), and that activity in superior temporal gyrus, and inferionr frontal gyrus, is sensitive to acquired community structure (Schapiro, Rogers, Cordova, Turk-Browne, & Botvinick, 2013). Participants will be divided into three groups, Transition 1, Transition 2 and

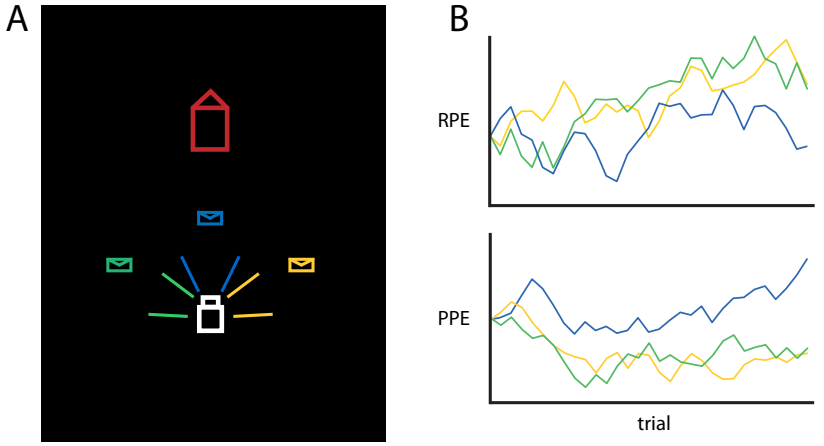

Figure 5.2. Task exploring the role of PPEs and RPEs in learning. (A) In this task participants can choose between three subgoals, highlighted in different colors. (B) Each subgoal is characterized by independent drifting functions of RPEs and PPEs.

Control (Figure 5.3A). They will be exposed to this statistical structure implicitly, performing an orthogonal cover task such as identification of a target.

Our main goals are to observe PPEs dependent on task structure (Figure 5.3B), and determine whether there is a relationship between the way participants encode the task structure and the degree of abstraction in PPEs. This would be done by correlating a mixture parameter (w) with similarity scores between stimuli within and across putative subtasks (Figure 5.3D) from activity in the temporal and frontal lobes (Schapiro et al., 2013, Figure 5.3D), during the Exposure Phase.

The Reward Phase of the task is very similar to the hierarchical task presented in Diuk et al. (2013). Participants have to choose between two casinos (Pillared houses in Figure 5.3B). Each casino is characterized by a distribution of "points", and a sequence of fractals. Upon choice, the amount of points necessary to leave the casino with additional money, is shown by a

dashed bar (Figure 5.3B). Points are earned by observing the outcome of the fractal bandits (lower level sequence of fractals in Figure 5.3B). If the amount of points is not attained, they participants leave the casino losing a certain amount. Each fractal is associated with a drifting probability of points, shown in Figure 5.3C, thus allowing continuous learning. Participants have no control over the sequence that is shown, in contrast with Diuk et al.. Trials from the Exposure Phase will be repeated in the Reward Phase, so that the induced structure is not forgotten.

Ventral striatum should be sensitive to PPEs at level 1 for participants in all three groups. Responses for PPEs at level 2 should depend statistical relationships between the fractals which participants were exposed to. Participants in groups Transition 1 and 2 should show extended PPEs at the 2^{nd} and 3^{rd} fractals respectively, whereas the control group should show no second level PPEs. The extent to which neural PPEs reflect this mixture of simultaneous PPEs at different levels, coded in the w parameter, should be correlated at the population level with the similarity between these two fractals.

Triggering PPEs while imaging midbrain dopaminergic nuclei. The midbrain dopaminergic nuclei are notoriously difficult to image, due to proximity to the basilar artery and susceptibility to cardiac interference (D'Ardenne et al., 2008). For this reason, it would be interesting to repeat the tasks presented in this thesis, and the one used in Diuk et al. (2013), while using methods appropriate for brainstem imaging.

Assessing preSMA activity in option policies and subsequent transfer. Single-neuron recordings in preSMA have shown neural responses to code for particular sequences of behavior (Shima et al., 1996; Nakamura et al., 1998; Shima & Tanji, 2000; Bor et al., 2003; Kennerley et al., 2004; Averbeck & Lee, 2007; Shima et al., 2007). According to model-free options, task structure should determined that certain states are associated with pseudo-reward, and thus independently reinforced.

The experimental group will be exposed to an MDP (MDP 1e, Figure 5.4A) with reward at one point and a perceptually salient intermediate state (subgoal). The MDP consists of states (images such as the fractals before), connected through arbitrary key pressings. If participants in this group treat the perceptually salient state as subgoal, and associate pseudo-reward with its attainment, then the policy leading to it (π_o) should be independently reinforced (Figure 5.4A, and see the section on subgoal discovery). In contrast, the control group exposed to MDP 1c, which has no parsing cues, should show no independent learning of a subpolicy. In a second session, both groups play in MDP 2 which only shares the subset of states leading to the subgoal (Figure 5.4B). We will assess how fast the experimental group learns the optimal policy (Figure 5.4C), which includes the previously learned option policy, compared with the control group. In addition, acquisition of option cached values predicts that the experimental group should show resistance to local devaluation (Figure 5.4C, "Devaluation"). We will perform this by introducing higher moving costs at certain transitions. In a single experiment we thus show positive and negative transfer effects of option policies.

Neurally, we posit that the extent to which an individual participant shows transfer of option policies should depend of the robustness of neural responses in preSMA. In addition, it should correlate with the caudality of striatal responses (Yin & Knowlton, 2006; Tricomi, Balleine, & ODoherty, 2009), consistent with a shift to habitual responses.

General directions

There is more in HRL than option-specific policies, PPEs, and extended PEs (see Figure 5.5A for a recapitulation of the neural mappings proposed for HRL). Future explorations of neural HRL should include tests of option identification, likely to reside in DLPFC (o in Figure 5.5A), explorations of the origin of pseudo-reward, $R_o(s)$, observations of the influence of task

structure in top-level values, and identification of the neural correlates of option-specific values, $V_o(s)$. In addition to these tests, it would be interesting to test whether the neural substrates for options are shared with those of state abstraction (e.g., Badre et al., 2010), even though the *options* framework usually does not employ state abstraction.

Further functional resolution can be achieved by manipulating the putative substrates of HRL. On this note, studies with functional disconnection (e.g., Parkinson et al., 2000), neuronal lesion (Yin, Knowlton, & Balleine, 2004), stimulation (Witten et al., 2011), and neural representation (Mulder, Nordquist, Örgüt, & Pennartz, 2003) would be decisive for the study of options. To our knowledge there are no animal studies directly addressing HRL hypotheses.

5.3 Comparison with Other Relevant Proposals

In this section we compare the neural HRL framework with other recent RL and non-RL accounts of hierarchical behavior.

5.3.1 Other RL models of hierarchical behavior

Haruno and Kawato (2006). The authors put forth a model of state abstraction along corticostriatal loops. In addition, they specify that activity should shift caudally in the course of learning (as in accounts of habitual learning, e.g., Tricomi et al., 2009), that difficulty of task dictates the loop that starts learning — more difficult tasks require more anterior loops, and that the posterior loops receive information from anterior loops.

In the model each loop keeps a specific Q value. The anterior PFC-basal loop computes an RPE using regular TD. The prediction error at the posterior loop is a weighted sum of the RPE of the anterior loop and a local RPE. Thus the term heterarchical. They present fMRI evidence for

the shift from caudate to putamen throughout learning, and confirm teh weighted PEs throughout the striatum.

Frank and Badre (2012). The authors offer a brain-based account of rule representation, according to RL. The model learns rules (stimuli-button press mappings) of a task presented in Badre et al. (2010). Stimuli have four dimensions and depending on the condition (hierarchical or flat), all dimensions might be relevant (flat) or the value in one dimension instructs which other dimensions to pay attention to, or to ignore. There is then, in the hierarchical condition, the potential for abstraction, i.e., for removing irrelevant dimensions. The authors provide a mechanistic account of previous neuroimaging data, showing segregation of abstract representations in DLPC (Badre et al., 2010), and correlate model behavior with individual performance in the task.

The model is a variant of the Prefrontal Basal Ganglia Working Memory model (PBWM, O'Reilly & Frank, 2006). This is a connectionist architecture where layers are connected according to the pattern of connectivity between PFC, midbrain and striatum. In Frank and Badre, the number of layers in PFC is extended to incorporate the rostrocaudal hypothesis of LPFC according to abstraction (Badre & D'Esposito, 2007).

The contribution of this paper was to show how abstract mappings can arise in PFC, without previously specifying them, drawing neuroimaging, behavioral and theoretical approaches. One caveat is that the architecture of the task is drawn into the model — "PFC" layer has the exact same number of units as possible abstract dimensions in the task (compare with Botvinick & Plaut, 2004, where no division of labor happens in hidden units).

Notable differences with options are the focus on state, instead of temporal, abstraction — though they might be both subserved by PFC and striatal loops; the use of a single dopaminergic signal for all levels of abstraction — dopamine gates relevant information into working memory — instead

of several independent teaching signals as posited in this thesis (RPEs and PPEs).

Ito and Doya (2011). This is similar to the previous accounts in that a division of labor is proposed along corticostriatal loops according to different levels of abstraction. In general, this is a proposal that resonates with *options*, though the link is not formally made. They posit that multiple Q values are learned at different levels of abstraction, simultaneously. In this framework, values of different abstraction are mapped onto the striatum: dorsolateral — more primitive, medioventral — more abstract. Curiously, this is the opposite of what is proposed by Bornstein and Daw (2011). Though they mention HRL, this is not a computational account, and the purpose is to propose an allocation of temporally abstract learning in corticostriatal loops.

Holroyd and Yeung (2011), and Holroyd and Yeung (2012). This proposal advances a theory of dorsal ACC in maintenance of temporally-extended behaviors. They leverage the fact that current theories cannot explain the effects of ACC lesion, producing, in the extreme, akinetic mutism, and at the same time predict the findings of other theories. It puts together hierarchical RL, research on cognitive control and human and animal lesions of dACC.

According to the theory, ACC represents the Q value of an option, which include: control costs of maintaining the option (which would not be present if an agent acted habitually and flat) and positive rewards accrued throughout the task (resonating with Shenhav et al., 2013). In case of a lesion, there is no representation of the overall benefit of an option, and action selection is based solely on local costs. Based on the Q value, ACC then guides DLPFC, and afterwards DLS, for implementing the option-specific policy. Recently, the authors have simulated the effect of ACC lesions and compared performance with other theories (Holroyd & McClure, submitted).

5.3.2 Non-RL models of hierarchical behavior

Representing hierarchical behavior, symbolic vs connectionist models. This distinction has mostly figured in preRL research of hierarchical behavioral (see Frank & Badre, 2012, for proposing a connectionist model of hierarchical RL), even though it is also relevant in HRL models, particularly in exploration of neural correlates. This is because, as pointed in Uithol et al. (2012), the action hierarchy might not have parallels in the control hierarchy.

In symbolic models (e.g., Miller et al., 1960; Estes, 1972; Norman & Shallice, 1986; Cooper & Shallice, 2000; Koechlin et al., 2003; Crump & Logan, 2010), there is a one-to-one mapping between control units and an action effect (stir in coffee, go to doctor, ...) and often the relationship between actions is built in. These models are important in describing aspects of behavior. However, they offer no account of how the units are learned, may import too many assumptions on the relationship between units (though for options to be useful, certain relationships between options must be prelearned, and similarly in MAXQ), and there might be no part of the brain with an activity mirroring the activity of the unit (Uithol et al., 2012, though this also applies to univariate exploration of RL correlates). In connectionist models (Elman, 1990, and Cleeremans, 1993, as cited in Botvinick, 2008; Botvinick and Plaut 2004; Botvinick 2007), there is no explicit, prelearned, division of labor between units. Rather, functions such as representing the identity of an option arise from the interaction between units. The contrast has been clearly made in Botvinick and Plaut (2004), where the same type of behavior as in Cooper and Shallice (2000) was modeled, without assuming a hierarchy of task units. In this thesis, we review Cooper and Shallice (2000), Botvinick and Plaut (2004), and Logan (2011).

Cooper and Shallice (2000). This builds upon the theory of action selection of Norman and Shallice (1986). The model focused on the scheduling of control units such that subtasks can be performed without

conflicting with parallel subtasks. When the inputs to a particular unit exceed a threshold, that unit is activated, and all other competing subtasks are inhibited. The relationship between units is built in into the model.

The authors successfully model routine behavior, including slips of action (Reason, 1979), and behavioral deficits following neurological damage, such as the Action Disorganization Syndrome (Schwartz et al., 1995).

Botvinick and Plaut (2004). The authors use a connectionist approach to model routine behavior, more specifically the production of coffee and tea. The model has three layers: *input*, a recurrent *hidden* layer, and an output layer. The input layer includes information about the current state of world objects, and attention to objects and previous actions (people act on what they are attending to). Output includes manipulative and perceptual actions (attend to object X). Learning happens by back-propagation of weights after feedback, in a way proportional to the contribution of the unit to the outcome.

The model was able to capture the similarity of actions in different contexts (e.g., pour sugar in tea vs pour sugar during coffee production). This similarity metric is not a feature of the standard *options* framework, which might allow for a parametric modulation of transfer. As in Cooper and Shallice (2000), the model produces regular and pathological slips of action. In addition it provides an account of learning and a mechanism for flexibility (a waiter has to adapt actions according to the costumer: one customer likes no sugar, another likes one scoop and another likes two scoops).

Logan (2011). This article reviews research (Crump & Logan, 2010, 2010) that uses typewriting as an exemplar domain where hierarchical properties of behavior can be tested. In order to predict the dynamics of typewriting, the authors stipulate a model with two levels of abstraction, termed outer (word level) and inner (keystroke level) loops. Besides temporal abstraction, the model also incorporates state abstraction: the information on

each level is encapsulated (the upper level does not care in which state the lower level is, as long as it is not in the state of completion). In addition, each level receives different forms of perceptual feedback.

The authors present behavioral evidence for interference effects at different levels of abstraction. Using Stroop-like tasks with typing color, it was shown that congruency affects RT but not interstroke interval: suggesting that interference occurs at level of words and not at keystrokes. In addition, scrambled sentences are typed as fast as normal sentences, but not words with scrambled letters. Another experiment showed that priming with a word benefits future writing of the first letter, but not of other letters in the word. Moreover, keystroke pressing relies on feel of the keyboard, whereas word production relies on visual spatial cues on the screen.

5.4 The Problem of Subgoal Discovery

Throughout this thesis we assumed that subgoals, related reward functions, and option policies, would be provided. While this encompasses a set of interesting learning problems by itself, it leaves open the question where do the options come from? This can be parsed into two questions, how are subgoals provided and where do the option policies come from? In this section we will focus on the first question, given that once subgoals are set, regular, model-free RL methods can be employed.

This is an important question as a particular subgoal can have a detrimental or beneficial effect on learning, compared to a flat agent. In addition, regardless of the usefulness of a single subgoal, adding options by itself increases the space of possible actions, something that was recognized in early AI as an *utility problem* (Lehman et al., 1996). As shown in Figure 5.6 in the rooms domain, the addition of options to the set of permissible actions has opposing effects of learning time, compared with a flat agent, depending on which subgoal the agent is given (see Botvinick, Niv, & Barto, 2009, and Jong et al., 2008, for more examples of negative and positive transfer).

With the exception of "fixed action patterns" — innate stereotypical sequences, which run to completion, resembling open-loop policies (Lorenz, 1950) — hierarchical behavior is acquired during development through accretion of subtasks (Bruner, 1973; Fischer, 1980, though "fixed action patterns" can serve as basis for later adaptive behavior, Thelen, 1981), or it can be acquired through a direct analysis of the learning problem (e.g., Solway et al., submitted, though the later is more aimed at providing an upper bound on the best subgoal partitioning).[1] Much like this distinction, in the computational literature some methods rely on direct analyses of a learning problem (e.g., Şimşek, Wolfe, & Barto, 2005), while others accrue options through experience (e.g., Bernstein, 1999).

[1] State and policy abstraction also seem to follow an increasing pattern throughout development (Halford, Wilson, & Phillips, 1998; Bunge & Zelazo, 2006).

In addition, approaches differ in the main regularity that is leveraged to identify useful subgoals. One set of approaches identifies candidate subgoals based on regularities of the state space: using graph-theoretical measures (Menache, Mannor, & Shimkin, 2002; Mannor, Menache, Hoze, & Klein, 2004; Şimşek et al., 2005; Solway et al., submitted, the latter also uses regularities in accrued rewards), unpredictability of transitions (Hengst, 2002), successful trajectories through certain states (Digney, 1998; McGovern & Barto, 2001), relative novelty of certain parts of the state space (Şimşek & Barto, 2004, separable from unpredictability, but still related to frequency of experience, in that it relies on building options to regions of the state space that are different from the ones the agent usually experiences), or based on salient perceptual properties of certain states (Singh et al., 2005, though this also relies on establishing a reward function). Other methods rely on statistics of policies, at individual (Thrun & Schwartz, 1995; Bernstein, 1999), or evolutionary timescales (Elfwing, Uchibe, Doya, & Christensen, 2007), or hierarchical imitation of policies (Friesen & Rao, 2010). Another family of approaches adds options by incrementally changing the reward function, by shaping (Kakade & Dayan, 2002, externally provided bonuses, much like parenting), or by setting as subgoals the initiation states of known policies (known as skill chaining Konidaris & Barto, 2009). Given the several angles from which an MDP can be carved, there are several attempts to formalize option creation as an optimization problem (Thrun & Schwartz, 1995; Foster & Dayan, 2002; Solway et al., submitted).

To our knowledge, with few exceptions (Kakade & Dayan, 2002; Reynolds, Zacks, & Braver, 2007; Solway et al., submitted), these principles have not been formally tested in psychology. Solway et al. (submitted) have shown that adult humans are optimal in their parsing of a task with regard to maximizing reward over the possible trajectories that can happen in an MDP.

Research on the neural basis of the acquisition of subgoals is still nascent (though, for comparison, there are neuroscientific accounts of rule-based learning in development, Bunge & Zelazo, 2006). Exploring the neural basis of subgoal discovery has yielded different structures dependent on the principle behind acquisition: analyses based on the state space might involve the temporal lobe (Schapiro et al., 2013) and likely the nucleus accumbens, given the connections with hippocampus (Haber & Knutson, 2010), policy-based methods might require dorsal striatal or lateral PFC engagement (Cole, Etzel, Zacks, Schneider, & Braver, 2011), and methods of intrinsic motivation might rely on the novelty properties of the dopaminergic system (Reed et al., 1996; Dayan & Balleine, 2002; Kakade & Dayan, 2002).

5.5 Model-based *Options*

The parallel between the model-free/habitual, model-based/goal-directed also extends to *options* (Balleine & Dickinson, 1998; Daw et al., 2005). The knowledge of the transition and reward functions is now used to make hierarchical predictions. In model-based *options* (Diuk, Strehl, & Littman, 2006; Jong & Stone, 2008), an agent skips over the transitions of primitive actions, to instead make temporally distant predictions — which state will it be at the end of the option, and what is its value. This resonates with the nature of planning in humans, using the example in Botvinick, Niv, and Barto (2009), "Perhaps I should buy one of those new cell phones... Well, that would cost me a few hundred dollars... But if I bought one, I could use it to check my email...". Computationally, this has the advantage of decreasing the size the search tree (Hayes-Roth, & Hayes-Roth, 1989; Kambhampati, Mali, & Srivastava, 1998; Marthi, Russell, & Wolfe, 2007, as cited in Botvinick, Niv, & Barto, 2009).

The evidence that humans make such extended predictions in the domain of action selection is rare, though not without precedents (Solway et

al., submitted; Huys et al., 2013). Besides the isolated examples, research on event perception (for a review, see Zacks et al., 2007) and action understanding (Mechsner, Kerzel, Knoblich, & Prinz, 2001; Hommel, Müsseler, Aschersleben, & Prinz, 2001) has shown that humans make predictions that reflect the task structure or the final goal, instead of a focus on the immediate consequences. Neurally, there is evidence that such extended predictions are encoded in the parietal cortex (Hamilton & Grafton, 2006, 2008).

A recent proposal and behavioral study (Dezfouli & Balleine, 2013) has integrated the two modes of control, model-free and model-based, with levels of abstraction. According to the proposal, behavior is best explained by a composition of habitual sequences — akin to option-specific policies — which are integrated by a model-based controller. Interestingly, the localization of habitual and goal-directed regions of the striatum (Bornstein & Daw, 2011) interacts with proposed striatal (Ito & Doya, 2011), and cortico-striatal (Badre, 2008; Frank & Badre, 2012) divisions of labor based on abstraction.

5.6 The Limits of the Hierarchy

Throughout this thesis we have assumed a strict hierarchy building upon primitive actions. However, RL methods can also be applied to continuous states (van Hasselt, 2012). Given that behavior is a sequence of muscle contractions (Hamilton & Grafton, 2007), and in principle RL mechanisms can bypass discretization of actions, what is the evidence for primitive actions?

Psychologically, humans do parse actions, and can do so at multiple levels of granularity (Schwartz et al., 1991; Zacks & Tversky, 2001), and it is known that infants and adults are sensitive to such structure and parse behavioral streams of actions in meaningful sequences (Reed, Montgomery, Schwartz, Palmer, & Pittenger, 1992; Zacks & Tversky, 2001; Baldwin, Baird, Saylor, & Clark, 2001), at different levels of abstraction. There is also

evidence for a neural hierarchical correspondence of control, from muscle properties, to kynematics, and finally to goals (Lemon et al., 1998, and, as cited in Hamilton & Grafton, 2007, Jackson, 1889, and Sherrington, 1906). It is interesting that evidence for an involvement of dopamine in learning specific visuomotor mappings is mixed (Weiner, Hallett, & Funkenstein, 1983; Contreras-Vidal & Buch, 2003; Isaias et al., 2011), sometimes finding a sparing of visuomotor adaptation in Parkinson's patients. Should the learning of specific kynematic and muscle properties not be dependent on dopamine, it posits the possibility that the "primitive actions" stipulated in most studies of neural RL, such as press left to select a bandit, are actually the ground level for dopamine mechanisms.

In theory, however, learning does not need to hinge on primitive actions, or continuous muscle properties. Some HRL methods, such as MAXQ or Feudal RL (Dayan & Hinton, 1993; Dietterich, 1998), can directly learn at a particular level independently of the bottom level. Though this can happen in *options*, there are limiting factors such as the fact that the state space is the same as the core MDP, and that it is not straightforward to represent policies that are not fully specified up to primitive actions, and instead only learn to call other options. An *options* agent (at least in the rigid formulation we have been discussing), learning to make a sandwich, would optimize the existing subtasks, such as "put lettuce", "open bread", and "put cheese", independently of making the overall sandwich, but would have a harder time learning that what actually matters is to get a combination of some bread, with some sort of vegetable plus some cheese or meat. More flexible models of human learning have been put forth, where agents learn a probability distribution over intermediate parts of the hierarchy (Wingate, Diuk, O'Donnell, Tenenbaum, & Gershman, 2013), which can exploit more sophisticated knowledge structures (e.g, "a sandwich is the same category as a wrap", Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010). Indeed,

there is some recent evidence that humans can learn directly at intermediate levels of the hierarchy (Huys et al., 2013).

Human tasks vary widely in the degree of abstraction (Barker & Wright, 1954). Even though highly temporally-extended tasks might have obvious hierarchical structure, it might not be useful anymore to exploit hierarchical structure in the same way it is stipulated in the *options* framework. Indeed, it might not be useful to build the subtask "get a PhD", as it is unlikely it will ever be transferred. On a speculative note, at this point, the interaction with generalization or the capacity for dynamically setting abstraction might come into play. Such cap on the current theories of behavior is also reflected in theories of neural representation of abstraction, which fail to offer an account for how PFC represents highly abstract behavior (Fuster, 1997; Badre, 2008).

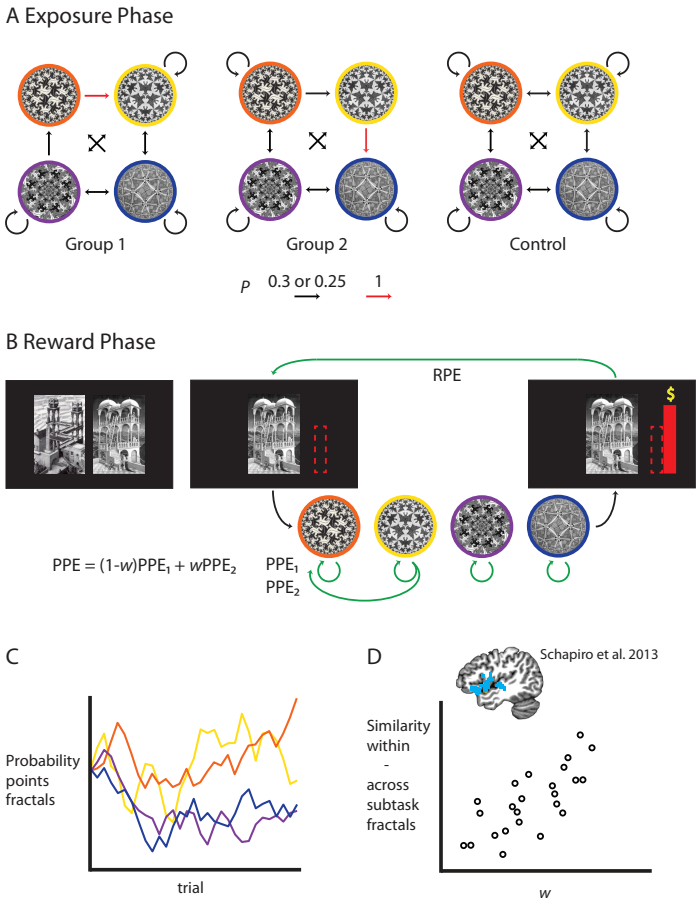

Figure 5.3. Manipulation of task structure independently of PPEs. (A) In this experiment participants are exposed to different statistical structures (Exposure Phase). (B) This structure will later be used to test whether neural PPEs in VS and dACC (Reward Phase) reflect a weighted sum of PPEs at two levels of abstraction. (C) The probabilities of points yielded by the fractals drifts according to a random walk in order to encourage learning. (D) Test whether at the population level (each dot is a hypothetical participant) similarity structure in superior temporal and inferior frontal gyri, as found in Schapiro et al. (2013), correlate with the weight of abstraction in PPEs.

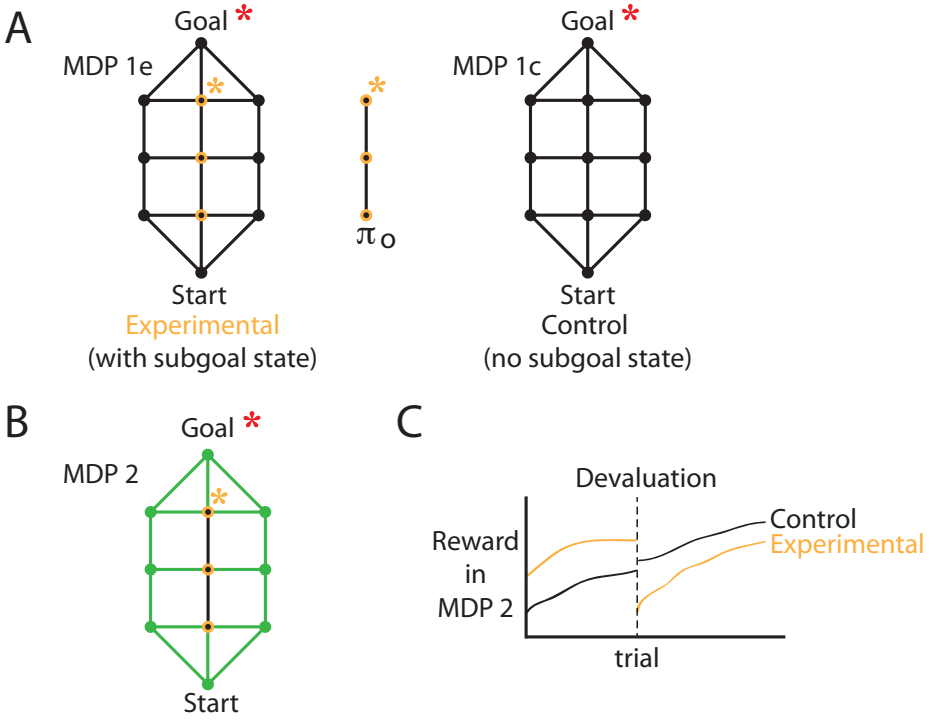

Figure 5.4. Positive and negative transfer of options. (A) Two groups of participants play a sequential key press task. In MPD 1e, one of the states is made to be perceptually salient (subgoal state marked by yellow asterisk, e.g., by playing a tone), whereas in MDP 1c all states are equally salient. Once participants have acquired proficiency in this task, they are transferred to MDP 2. (B) MDP 2 shares an intermediate set of states with MDP 1e, and none with MDP 1c. (C) Performance in MDP 2. The group trained in MDP 1e should outperform the group trained in MDP 1c. However, upon devaluation of any of the state-action pairs in π_o, the control group should show faster avoidance of the devalued sequence.

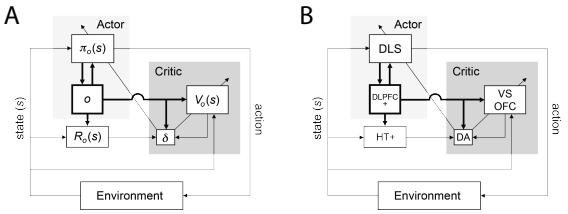

Figure 5.5. Actor-critic implementation of HRL and proposed neural extensions.

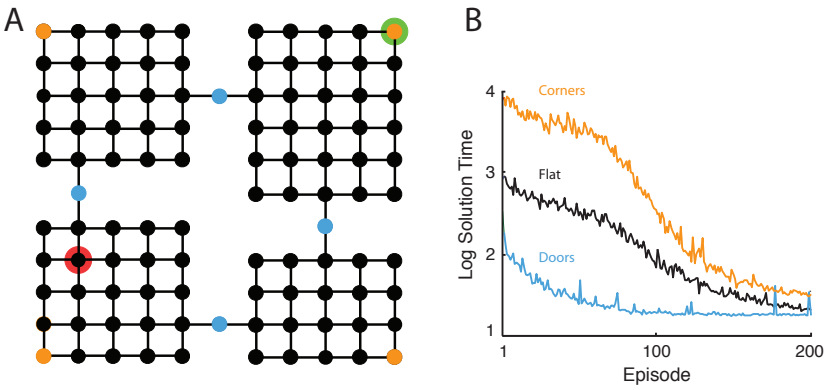

Figure 5.6. Examples of positive and negative transfer in the rooms domain. (A) Consider the rooms gridworld, where a room is composed of a series of states, each depicted by a black dot and admissible transitions with links between the states (green = start, red = goal, orange = corner states, blue = door/subgoal states). Transitions between rooms happen through a single state, which we call doors. One HRL agent was endowed with a "get-to-door" option, another was endowed with a "get-to-corner-of-room" option, and both agents had a primitive actions in their behavioral repertoires. A third agent, "flat" agent only had primitive actions. (B) The option "get-to-door" decreased the time required to reach an optimal policy compared with the flat agent, whereas "get-to-corner-of-room" delayed the attainment of an optimal policy. Adapted with permission from Solway et al. (submitted).

References

Aldridge, J. W., & Berridge, K. C. (1998). Coding of serial order by neostriatal neurons: a "natural action" approach to movement sequence. *Journal of Neuroscience*, *18*(7), 2777-2787.

Aldridge, J. W., Berridge, K. C., Herman, M., & Zimmer, L. (1993). Neuronal coding of serial order: syntax of grooming in the neostriatum. *Psychological Science*, *4*(6), 391–395.

Alexander, G. E., DeLong, M. R., & Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, *9*, 357-381.

Alexander, W. H., & Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, *14*(10), 1338–1344.

Amiez, C., Joseph, J., & Procyk, E. (2005). Anterior cingulate error-related activity is modulated by predicted reward. *European Journal of Neuroscience*, *21*(12), 3447-3452.

Amiez, C., & Petrides, M. (2007). Selective involvement of the mid-dorsolateral prefrontal cortex in the coding of the serial order of visual stimuli in working memory. *Proceedings of the National Academy of Sciences*, *104*(34), 13786–13791.

Asaad, W. F., Rainer, G., & Miller, E. K. (2000). Task-specific neural activity in the primate prefrontal cortex. *Journal of Neurophysiology*, *84*(1), 451–459.

Averbeck, B. B., & Lee, D. (2007). Prefrontal neural correlates of memory for sequences. *Journal of Neuroscience*, *27*(9), 2204-2211.

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

Badre, D., & D'Esposito, M. (2007). Functional magnetic resonance imaging evidence for a hierarchical organization of the prefrontal cortex. *Journal of Cognitive Neuroscience*, *19*(12), 2082–2099.

Badre, D., & D'Esposito, M. (2009). Is the rostro-caudal axis of the frontal lobe hierarchical? *Nature Reviews Neuroscience*, *10*(9), 659–669.

Badre, D., Hoffman, J., Cooney, J. W., & D'Esposito, M. (2009). Hierarchical cognitive control deficits following damage to the human frontal lobe. *Nature Neuroscience*, *12*(4), 515–522.

Badre, D., Kayser, A. S., & D Esposito, M. (2010). Frontal Cortex and the Discovery of Abstract Action Rules. *Neuron*, *66*(2), 315–326.

Baker, T. E., & Holroyd, C. B. (2011). Dissociated roles of the anterior cingulate cortex in reward and conflict processing as revealed by the feedback error-related negativity and N200. *Biological Psychology*, *87*(1), 25-34.

Baldwin, D. A., Baird, J. A., Saylor, M. M., & Clark, M. A. (2001). Infants parse dynamic action. *Child development*, *72*(3), 708–717.

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37*(4), 407-419.

Barker, R., & Wright, H. F. (1954). Dividing the behavior stream. In *Midwest and its children: The psychological ecology of an american town.* New York, NY: Row, Peterson and Company.

Barnes, T. D., Mao, J.-B., Hu, D., Kubota, Y., Dreyer, A. A., Stamoulis, C., ... Graybiel, A. M. (2011). Advance cueing produces enhanced action-boundary patterns of spike activity in the sensorimotor striatum. *Journal of Neurophysiology*, *105*(4), 1861–1878.

Barto, A. G. (1995). Adaptive critics and the basal ganglia. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (p. 215-232). Cambridge, MA: MIT Press.

Barto, A. G., & Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems: Theory and Applications*, *13*(4), 341-379.

Barto, A. G., & Sutton, R. S. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, *88*(2), 135-170.

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike adaptive

elements that can solve difficult learning control problems. *Systems, Man and Cybernetics, IEEE Transactions on*(5), 834–846.

Behrmann, M., & Tipper, S. P. (1999). Attention accesses multiple reference frames: evidence from visual neglect. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(1), 83-101.

Bellman, R. (1957). *Dynamic programming.* Princeton, NJ: Princeton University Press.

Berdyyeva, T. K., & Olson, C. R. (2010). Rank signals in four areas of macaque frontal cortex during selection of actions and objects in serial order. *Journal of Neurophysiology*, *104*(1), 141-159.

Berendse, H. W., Graaf, Y. G.-D., & Groenewegen, H. J. (1992). Topographical organization and relationship with ventral striatal compartments of prefrontal corticostriatal projections in the rat. *Journal of Comparative Neurology*, *316*(3), 314–347.

Bernstein, D. S. (1999). *Reusing old policies to accelerate learning on new MDPs* (Tech. Rep. No. 99-26). University of Massachusetts, Amherst.

Berridge, K. C., Fentress, J. C., & Parr, H. (1987). Natural syntax rules control action sequence of rats. *Behavioural Brain Research*, *23*(1), 59-68.

Bertsekas, D. P., & Tsitsiklis, J. (1996). *Neuro-dynamic programming.* Belmont, MA: Athena Scientific.

Bonini, L., Serventi, F. U., Simone, L., Rozzi, S., Ferrari, P. F., & Fogassi, L. (2011). Grasping neurons of monkey parietal and premotor cortices encode action goals at distinct levels of abstraction during complex action sequences. *Journal of Neuroscience*, *31*(15), 5876–5886.

Bor, D., Duncan, J., Wiseman, R. J., & Owen, A. M. (2003). Encoding strategies dissociate prefrontal activity from working memory demand. *Neuron*, *37*(2), 361-367.

Bornstein, A. M., & Daw, N. D. (2011). Multiplicity of control in the basal ganglia: computational roles of striatal subregions. *Current Opinion in Neurobiology*, *21*(3), 374–380.

Botvinick, M., Nystrom, L. E., Fissell, K., Carter, C. S., & Cohen, J. D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature*, *402*(6758), 179-181.

Botvinick, M., & Plaut, D. C. (2004). Doing without schema hierarchies: a recurrent connectionist approach to normal and impaired routine

sequential action. *Psychological Review*, *111*(2), 395-429.

Botvinick, M. M. (2007). Multilevel structure in behaviour and in the brain: a model of Fuster's hierarchy. *Philosophical Transactions of the Royal Society of London Series B, Biological sciences*, *362*(1485), 1615–1626.

Botvinick, M. M. (2008). Hierarchical models of behavior and prefrontal function. *Trends in Cognitive Sciences*, *12*(5), 201-208.

Botvinick, M. M., Braver, T. S., Barch, D., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652.

Botvinick, M. M., Huffstetler, S., & Mcguire, J. T. (2009). Effort discounting in human nucleus accumbens. *Cognitive, Affective, & Behavioral Neuroscience*, *9*(1), 16–27.

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

Boyd, L., Edwards, J., Siengsukon, C., Vidoni, E., Wessel, B., & Linsdell, M. (2009). Motor sequence chunking is impaired by basal ganglia stroke. *Neurobiology of Learning and Memory*, *92*(1), 35–44.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436.

Breiter, H. C., Aharon, I., Kahneman, D., Dale, A., & Shizgal, P. (2001). Functional imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron*, *30*(2), 619–639.

Bruner, J. S. (1973). Organization of early skilled action. *Child Development*, *44*(1), 1–11.

Bunge, S. A. (2004). How we use rules to select actions: a review of evidence from cognitive neuroscience. *Cognitive, Affective, & Behavioral Neuroscience*, *4*(4), 564–579.

Bunge, S. A., & Zelazo, P. D. (2006). A brain-based account of the development of rule use in childhood. *Current Directions in Psychological Science*, *15*(3), 118–121.

Burkhardt, J. M., Jin, X., & Costa, R. M. (2009). Dissociable effects of dopamine on neuronal firing rate and synchrony in the dorsal striatum. *Frontiers in Integrative Neuroscience*, *3*(28).

Christoff, K. (2003). Evaluating self-generated information: anterior pre-

frontal contributions to human cognition. *Behavioral Neuroscience*, *117*(6), 1161-1168.

Christoff, K., & Keramatian, K. (2007). Abstraction of mental representations: theoretical considerations and neuroscientific evidence. In S. A. Bunge & J. D. Wallis (Eds.), *Perspectives on rlue-guided behavior* (p. 107-126). Oxford: Oxford University Press.

Christoph, G. R., Leonzio, R. J., & Wilcox, K. S. (1986). Stimulation of the lateral habenula inhibits dopamine-containing neurons in the substantia nigra and ventral tegmental area of the rat. *Journal of Neuroscience*, *6*(3), 613–619.

Cole, M. W., Etzel, J. A., Zacks, J. M., Schneider, W., & Braver, T. S. (2011). Rapid transfer of abstract rules to novel contexts in human lateral prefrontal cortex. *Frontiers in human neuroscience*, *5*(142).

Collins, A. G., & Frank, M. J. (2013). Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychological Review*, *120*(1), 190–229.

Contreras-Vidal, J. L., & Buch, E. R. (2003). Effects of parkinson's disease on visuomotor adaptation. *Experimental brain research*, *150*(1), 25–32.

Conway, C. M., & Christiansen, M. H. (2001). Sequential learning in non-human primates. *Trends in Cognitive Sciences*, *5*(12), 539–546.

Cooper, J. C., & Knutson, B. (2008). Valence and salience contribute to nucleus accumbens activation. *Neuroimage*, *39*(1), 538–547.

Cooper, R., & Shallice, T. (2000). Contention scheduling and the control of routine activities. *Cognitive Neuropsychology*, *17*(4), 297–338.

Corbetta, M., Patel, G., & Shulman, G. L. (2008). The reorienting system of the human brain: from environment to theory of mind. *Neuron*, *58*(3), 306–324.

Courtney, S. M., Roth, J. K., & Sala, J. B. (2007). A hierarchical biased-competition model of domain-dependent working memory maintenance and executive control. In N. Osaka, R. H. Logie, & M. D'Esposito (Eds.), *The cognitive neuroscience of working memory* (p. 369-399). Oxford: Oxford University Press.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, *10*(7), 294-300.

Cox, R. W. (1996). Afni: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research*, *29*(3), 162–173.

Cromwell, H. C., & Berridge, K. C. (1996). Implementation of action sequences by a neostriatal site: a lesion mapping study of grooming syntax. *Journal of Neuroscience*, *16*(10), 3444–3458.

Croxson, P. L., Johansen-Berg, H., Behrens, T. E., Robson, M. D., Pinsk, M. A., Gross, C. G., ... Rushworth, M. F. (2005). Quantitative investigation of connections of the prefrontal cortex in the human and macaque using probabilistic diffusion tractography. *Journal of Neuroscience*, *25*(39), 8854–8866.

Croxson, P. L., Walton, M. E., O'Reilly, J. X., Behrens, T. E., & Rushworth, M. F. (2009). Effort-based cost–benefit valuation and the human brain. *Journal of Neuroscience*, *29*(14), 4531–4541.

Crump, M. J., & Logan, G. D. (2010). Hierarchical control and skilled typing: evidence for word-level control over the execution of individual keystrokes. *Journal of Experimental Psychology: Learning Memory and Cognition*, *36*(6), 1369–1380.

Crump, M. J. C., & Logan, G. D. (2010). Episodic contributions to sequential control: learning from a typist's touch. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(3), 662–672.

D'Ardenne, K., McClure, S. M., Nystrom, L. E., & Cohen, J. D. (2008). BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. *Science*, *319*(5867), 1264–1267.

Daw, N. D., Courville, A. C., & Touretzky, D. S. (2003). Timing and partial observability in the dopamine system. In *Advances in neural information processing systems* (Vol. 15, p. 99-106). Cambridge, MA: MIT Press.

Daw, N. D., & Frank, M. J. (2009). Reinforcement learning and higher level cognition: introduction to special issue. *Cognition*, *113*(3), 259–261.

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

Daw, N. D., Niv, Y., & Dayan, P. (2006). Actions, policies, values and the basal ganglia. In E. Bezard (Ed.), *Recent breakthroughs in basal ganglia research* (p. 91-106). Hauppage, NY: Nova Science.

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

Dayan, P., & Hinton, G. (1993). Feudal reinforcement learning. In *Advances in neural information processing systems 5* (pp. 271–278).

Dayan, P., & Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. *Current Opinion in Neurobiology*, *18*(2), 185–196.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience*, *18*(1), 193–222.

D'Esposito, M. (2007). From cognitive to neural models of working memory. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *362*(1481), 761-772.

Dezfouli, A., & Balleine, B. W. (2013). Evidence that goal-directed and habitual action control are hierarchically organized. In *Reinforcement learning and decision making 2013.*

Dickinson, A. (1985). Actions and habits: The development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London. . . .*, *308*(1135), 67-78.

Dietterich, T. G. (1998). The MAXQ method for hierarchical reinforcement learning. In *Proceedings of the fifteenth international conference on machine learning* (Vol. 8, p. 118-126).

Digney, B. L. (1998). Learning hierarchical control structures for multiple tasks and changing environments. In *Proceedings of the fifth international conference on simulation of adaptive behavior on from animals to animats* (Vol. 5, pp. 321–330).

Diuk, C., & Littman, M. L. (2008). Hierarchical reinforcement learning. In J. R. R. Dopico, J. Dorado, & A. Pazos (Eds.), *Encyclopedia of artificial intelligence* (p. 825-830). Hershey, PA: IGI Global.

Diuk, C., Strehl, A. L., & Littman, M. L. (2006). A hierarchical approach to efficient reinforcement learning in deterministic domains. In *Proceedings of the fifth international joint conference on autonomous agents and multiagent systems* (pp. 313–319).

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*(13), 5797-5805.

Duncan, J. (2013). The structure of cognition: Attentional episodes in mind and brain. *Neuron*, *80*(1), 35–50.

Džeroski, S., De Raedt, L., & Driessens, K. (2001). Relational reinforcement learning. *Machine Learning*, *43*(1-2), 7–52.

Elfwing, S., Uchibe, E., Doya, K., & Christensen, H. I. (2007). Evolutionary development of hierarchical learning structures. *Evolutionary Computation, IEEE Transactions on*, *11*(2), 249–264.

Estes, W. K. (1972). An associative basis for coding and organization in memory. *Coding processes in human memory*, 161–190.

Farooqui, A. A., Mitchell, D., Thompson, R., & Duncan, J. (2012). Hierarchical organization of cognition reflected in distributed frontoparietal activity. *The Journal of Neuroscience*, *32*(48), 17373–17381.

Fentress, J. C. (1972). Development and patterning of movement sequences in inbred mice. In *The biology of behavior* (p. 83-132). Corvallis, OR: Oregon State University.

Fikes, R. E., Hart, P. E., & Nilsson, N. J. (1972). Learning and executing generalized robot plans. *Artificial Intelligence*, *3*, 251–288.

Fischer, K. W. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review*, *87*(6), 477-531.

Foster, D., & Dayan, P. (2002). Structure in the space of value functions. *Machine Learning*, *49*(2-3), 325–346.

Fountain, S. B., Wallace, D. G., & Rowan, J. D. (2002). *The organization of sequential behavior.* New York, NY: Springer.

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

Frank, M. J., & Claus, E. D. (2006). Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological Review*, *113*(2), 300–326.

Frank, M. J., & Seeberger, L. C. (2004). By carrot or by stick: Cognitive reinforcement learning in parkinsonism. *Science*, *306*(5703), 1940-1943.

Friesen, A. L., & Rao, R. P. N. (2010). Imitation learning with hierarchical actions. In *Development and learning (icdl), 2010 ieee 9th international conference on.* IEEE.

Fujii, N., & Graybiel, A. M. (2003). Representation of action sequence boundaries by macaque prefrontal cortical neurons. *Science*,

301(5637), 1246–1249.

Fujii, N., & Graybiel, A. M. (2005). Time-varying covariance of neural activities recorded in striatum and frontal cortex as monkeys perform sequential-saccade tasks. *Proceedings of the National Academy of Sciences*, *102*(25), 9032–9037.

Fuster, J. M. (1997). *The prefrontal cortex: anatomy, physiology, and neuropsychology of the frontal lobe*. Philadelphia, PA: Lippincott-Raven.

Fuster, J. M. (2001). The prefrontal cortex — an update: time is of the essence. *Neuron*, *30*(2), 319-333.

Garnham, A., Shillcock, R. C., Brown, G. D., Mill, A. I., & Cutler, A. (1981). Slips of the tongue in the london-lund corpus of spontaneous conversation. *Linguistics*, *19*(7-8), 805–818.

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science*, *4*(6), 385–390.

Gehring, W. J., & Willoughby, A. R. (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science*, *295*(5563), 2279-2282.

Gershman, S. J., & Niv, Y. (2012). Exploring a latent cause theory of classical conditioning. *Learning & Behavior*, *40*(3), 255-68.

Ghavamzadeh, M., & Mahadevan, S. (2001). Continuous-time hierarchical reinforcement learning. In *Proceedings of the eighteenth international conference on machine learning* (p. 186-193). Morgan Kauffmann.

Gilhousen, H. (1940). Final goal versus sub-goal distance discrimination. *Journal of Comparative Psychology*, *31*(1), 35–42.

Goldman-Rakic, P. S. (1987). Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. *Comprehensive Physiology*.

Grafman, J. (2002). The human prefrontal cortex has evolved to represent components of structured event complexes. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology: The frontal lobes* (Vol. 7). Amsterdam: Elsevier.

Graybiel, A. M. (1998). The basal ganglia and chunking of action repertoires. *Neurobiology of Learning and Memory*, *70*(1-2), 119–136.

Grent-'t Jong, T., & Woldorff, M. G. (2007). Timing and sequence of brain activity in top-down control of visual-spatial attention. *PLoS Biology*,

5(1), e12.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in cognitive sciences*, *14*(8), 357–364.

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

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences*, *21*(6), 803–831.

Hamilton, A., & Grafton, S. (2006). Goal Representation in Human Anterior Intraparietal Sulcus. *Journal of Neuroscience*, *26*(4), 1133-1137.

Hamilton, A. F. d. C., & Grafton, S. T. (2007). The motor hierarchy: from kinematics to goals and intentions. In Y. Rossetti, M. Kawato, & P. Haggard (Eds.), *Attention and performance* (Vol. 22). Cambridge, MA: MIT Press.

Hamilton, A. F. d. C., & Grafton, S. T. (2008). Action outcomes are represented in human inferior frontoparietal cortex. *Cerebral Cortex*, *18*(5), 1160–1168.

Hare, T., O'Doherty, J., Camerer, C. F., Schultz, W., & Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *Journal of Neuroscience*, *28*(22), 5623-5630.

Haruno, M., & Kawato, M. (2006). Heterarchical reinforcement-learning model for integration of multiple cortico-striatal loops: fMRI examination in stimulus-action-reward association learning. *Neural Networks*, *19*(8), 1242-1254.

Hasegawa, R. P., Blitz, A. M., & Goldberg, M. E. (2004). Neurons in monkey prefrontal cortex whose activity tracks the progress of a three-step self-ordered task. *Journal of Neurophysiology*, *92*(3), 1524–1535.

Hayden, B. Y., Heilbronner, S. R., Pearson, J. M., & Platt, M. L. (2011). Surprise signals in anterior cingulate cortex: neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. *Journal of Neuroscience*, *31*(11), 4178–4187.

Hengst, B. (2002). Discovering hierarchy in reinforcement learning with

hexq. In *Icml* (Vol. 2, pp. 243–250).

Hengst, B. (2012). Hierarchical approaches. In M. Wiering & M. Otterlo (Eds.), *Reinforcement learning* (p. 293-323). Berlin: Springer.

Hestvik, A., Maxfield, N., Schwartz, R. G., & Shafer, V. (2007). Brain responses to filled gaps. *Brain and Language*, *100*(3), 301–316.

Holroyd, C. B., & Coles, M. G. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*(4), 679–709.

Holroyd, C. B., & McClure, S. M. (submitted). *Hierarchical control over effortful behavior by anterior cingulate cortex.*

Holroyd, C. B., Nieuwenhuis, S., Yeung, N., & Cohen, J. D. (2003). Errors in reward prediction are reflected in the event-related brain potential. *NeuroReport*, *14*(18), 2481–2484.

Holroyd, C. B., Nieuwenhuis, S., Yeung, N., Nystrom, L. E., Mars, R. B., Coles, M. G., & Cohen, J. D. (2004). Dorsal anterior cingulate cortex shows fMRI response to internal and external error signals. *Nature Neuroscience*, *7*(5), 497–498.

Holroyd, C. B., & Yeung, N. (2011). An integrative theory of anterior cingulate cortex function: Option selection in hierarchical reinforcement learning. In R. B. Mars, J. Sallet, M. F. S. Rushworth, & N. Yeung (Eds.), *Neural Basis of Motivational and Cognitive Control - Rogier B. Mars - Google Books* (pp. 333–349). Cambridge, MA: MIT Press.

Holroyd, C. B., & Yeung, N. (2012). Motivation of extended behaviors by anterior cingulate cortex. *Trends in Cognitive Sciences*, *16*(2), 122-128.

Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences*, *24*(05), 849–878.

Hoshi, E. (2006). Functional specialization within the dorsolateral prefrontal cortex: a review of anatomical and physiological studies of non-human primates. *Neuroscience Research*, *54*(2), 73–84.

Hoshi, E., Shima, K., & Tanji, J. (1998). Task-dependent selectivity of movement-related neuronal activity in the primate prefrontal cortex. *Journal of Neurophysiology*, *80*(6), 3392–3397.

Houk, J. C., Adams, J. L., & Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement.

In J. C. Houk, J. Davis, & D. Beiser (Eds.), *Models of information processing in the basal ganglia* (p. 249-270). Cambridge, MA: MIT Press.

Humphreys, G. W., & Forde, E. M. E. (1998). Disordered action schemas and action disorganization syndrome. *Cognitive Neuropsychology*, *15*(6), 771-812.

Huys, Q., Lally, N., Falkner, P., Gershman, S., Dayan, P., & Roiser, J. (2013). Hierarchical deconstruction and memoization of goal-directed plans. In *Reinforcement learning and decision making 2013.*

Isaias, I. U., Moisello, C., Marotta, G., Schiavella, M., Canesi, M., Perfetti, B., ... Ghilardi, M. F. (2011). Dopaminergic striatal innervation predicts interlimb transfer of a visuomotor skill. *The Journal of Neuroscience*, *31*(41), 14458–14462.

Ito, M., & Doya, K. (2011). Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit. *Current Opinion in Neurobiology*, *21*(3), 368-373.

Jessup, R. K., Busemeyer, J. R., & Brown, J. W. (2010). Error effects in anterior cingulate cortex reverse when error likelihood is high. *Journal of Neuroscience*, *30*(9), 3467–3472.

Jin, X., & Costa, R. M. (2010). Start/stop signals emerge in nigrostriatal circuits during sequence learning. *Nature*, *466*(7305), 457–462.

Jocham, G., & Ullsperger, M. (2009). Neuropharmacology of performance monitoring. *Neuroscience & Biobehavioral Reviews*, *33*(1), 48–60.

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

Johnston, K., & Everling, S. (2006). Neural activity in monkey prefrontal cortex is modulated by task context and behavioral instruction during delayed-match-to-sample and conditional prosaccade–antisaccade tasks. *Journal of Cognitive Neuroscience*, *18*(5), 749-765.

Jong, N. K., Hester, T., & Stone, P. (2008). The utility of temporal abstraction in reinforcement learning. In *Proceedings of the 7th international joint conference on autonomous agents and multiagent systems-volume 1* (pp. 299–306).

Jong, N. K., & Stone, P. (2008). Hierarchical model-based reinforcement learning: R-max+ MAXQ. In *Proceedings of the 25th international*

conference on machine learning (pp. 432–439).

Kakade, S., & Dayan, P. (2002). Dopamine: generalization and bonuses. *Neural Networks*, *15*(4), 549–559.

Kalis, A., Kaiser, S., & Mojzisch, A. (2013). Why we should talk about option generation in decision-making research. *Frontiers in Psychology*, *4*(555).

Kendler, H. (1943). The influence of a sub-goal on maze behavior. *Journal of Comparative Psychology*, *36*(2), 67–73.

Kennerley, S. W., Sakai, K., & Rushworth, M. F. S. (2004). Organization of action sequences and the role of the pre-sma. *Journal of Neurophysiology*, *91*(2), 978–993.

Kermadi, I., & Joseph, J. (1995). Activity in the caudate nucleus of monkey during spatial sequencing. *Journal of Neurophysiology*, *74*(3), 911–933.

Kermadi, I., Jurquet, Y., Arzi, M., & Joseph, J. (1993). Neural activity in the caudate nucleus of monkeys during spatial sequencing. *Experimental Brain Research*, *94*(2), 352–356.

Klein, G., Wolf, S., Militello, L., & Zsambok, C. (1995). Characteristics of skilled option generation in chess. *Organizational Behavior and Human Decision Processes*, *62*(1), 63–69.

Knutson, B., Taylor, J., Kaufman, M., & Peterson, R. (2005). Distributed Neural Representation of Expected Value. *Journal of Neuroscience*, *25*(19), 4806-4812.

Knutson, B., Wood, J. N., & Grafman, J. (2004). Brain activation in processing temporal sequence: an fmri study. *Neuroimage*, *23*(4), 1299-1307.

Koechlin, E., Ody, C., & Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science*, *302*(5648), 1181–1185.

Koechlin, E., & Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences*, *11*(6), 229–235.

Konidaris, G., & Barto, A. G. (2009). Skill discovery in continuous reinforcement learning domains using skill chaining. In *Advances in neural information processing systems* (pp. 1015–1023).

Krebs, R. M., Boehler, C. N., Roberts, K. C., Song, A. W., & Woldorff,

M. G. (2012). The involvement of the dopaminergic midbrain and cortico-striatal-thalamic circuits in the integration of reward prospect and attentional task demands. *Cerebral Cortex*, *22*(3), 607–615.

Krigolson, O. E., & Holroyd, C. B. (2006). Evidence for hierarchical error processing in the human brain. *Neuroscience*, *137*(1), 13–17.

Krigolson, O. E., & Holroyd, C. B. (2007). Hierarchical error processing: different errors, different systems. *Brain Research*, *1155*, 70–80.

Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences*, *12*(2), 72-9.

Lashley, K. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral Mechanisms in Behavior.* New York, NY: John Wiley & Sons.

Lawson, R. P., Drevets, W. C., & Roiser, J. P. (2012). Defining the habenula in human neuroimaging studies. *NeuroImage*, *64*, 722-727.

Lehman, J. F., Laird, J. E., & Rosenbloom, P. (1996). A gentle introduction to Soar, an architecture for human cognition. In S. Scarborough & D. Sternberg (Eds.), *Invitation to cognitive science* (Vol. 4, pp. 212–249). MIT Press.

Lemon, R. N., Baker, S. N., Davis, J. A., Kirkwood, P. A., Maier, M. A., & Yang, H. S. (1998). The importance of the cortico-motoneuronal system for control of grasp. In *Novartis foundation symposium* (Vol. 218, pp. 202–215).

Li, L., Walsh, T. J., & Littman, M. L. (2006). Towards a unified theory of state abstraction for mdps. In *Proceedings of the ninth international symposium on artificial intelligence and mathematics.*

Logan, G. D. (2011). Hierarchical control of cognitive processes: The case for skilled typewriting. In B. H. Ross (Ed.), *Psychology of Learning and Motivation: Advances in Research and Theory* (Vol. 54, p. 1-28). San Diego, CA: Elsevier.

Lorenz, K. Z. (1950). The comparative method in studying innate behavior patterns. In *Physiological mechanisms in animal behavior. (Society's Symposium IV.)* (p. 221-268). Oxford: Academic Press.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*(4), 276-298.

MacLeod, C. M. (1991). Half a century of research on the stroop effect: an

integrative review. *Psychological Bulletin*, *109*(2), 163-203.

Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience*, *14*(2), 154–162.

Mannor, S., Menache, I., Hoze, A., & Klein, U. (2004). Dynamic abstraction in reinforcement learning via clustering. In *Proceedings of the twenty-first international conference on machine learning* (p. 71).

Mars, R. B., Coles, M. G. H., Grol, M. J., Holroyd, C. B., Nieuwenhuis, S., Hulstijn, W., & Toni, I. (2005). Neural dynamics of error processing in medial frontal cortex. *NeuroImage*, *28*(4), 1007–1013.

Matsumoto, M., & Hikosaka, O. (2007). Lateral habenula as a source of negative reward signals in dopamine neurons. *Nature*, *447*(7148), 1111–1115.

McGovern, A., & Barto, A. G. (2001). Automatic discovery of subgoals in reinforcement learning using diverse density. *Computer Science Department Faculty Publication Series*, 8.

Mechsner, F., Kerzel, D., Knoblich, G., & Prinz, W. (2001). Perceptual basis of bimanual coordination. *Nature*, *414*(6859), 69–73.

Menache, I., Mannor, S., & Shimkin, N. (2002). Q-cut—dynamic discovery of sub-goals in reinforcement learning. In *Proceedings of the thirteenth european conference on machine learning* (p. 295-306). Berlin, Heidelberg: Springer.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortical function. *Annual Review of Neuroscience*, *24*(1), 167–202.

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the Structure of Behavior*. New York, NY: Holt, Rinehart and Winston.

Miller, P. T. A., M. W. and, & Vogt, B. A. (2009). Dopamine systems in the cingulate gyrus: Organization, development and neurotoxic vulnerability. In B. A. Vogt (Ed.), *Cingulate neurobiology and disease.* Oxford: Oxford University Press.

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*(6), 788-798.

Miyamoto, H., Morimoto, J., Doya, K., & Kawato, M. (2004). Reinforcement learning with via-point representation. *Neural Networks*, *17*(3),

299-305.

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, *16*(5), 1936-1947.

Morris, G., Arkadir, D., Nevet, A., Vaadia, E., & Bergman, H. (2004). Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. *Neuron*, *43*(1), 133-143.

Mulder, A. B., Nordquist, R. E., Örgüt, O., & Pennartz, C. (2003). Learning-related changes in response patterns of prefrontal neurons during instrumental conditioning. *Behavioural Brain Research*, *146*(1), 77–88.

Mushiake, H., & Strick, P. L. (1995). Pallidal neuron activity during sequential arm movements. *Journal of Neurophysiology*, *74*(6), 2754–2758.

Nakamura, K., Sakai, K., & Hikosaka, O. (1998). Neuronal activity in medial frontal cortex during learning of sequential procedures. *Journal of Neurophysiology*, *80*(5), 2671–2687.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, *9*(3), 353–383.

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

Nieuwenhuis, S., Heslenfeld, D. J., von Geusau, N. J. A., Mars, R. B., Holroyd, C. B., & Yeung, N. (2005). Activity in human reward-sensitive brain areas is strongly context dependent. *Neuroimage*, *25*(4), 1302–1309.

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

Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.), *Consciousness and self-regulation: advances in research and theory* (p. 1-18). New York, NY: Plenum Press.

O'Doherty, J. P., Buchanan, T. W., Seymour, B., & Dolan, R. J. (2006). Predictive neural coding of reward preference involves dissociable responses in human ventral midbrain and ventral striatum. *Neuron*, *49*(1), 157-166.

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

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*(5669), 452–454.

O'Doherty, J. P., Hampton, A., & 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'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, *18*(2), 283–328.

Pagnoni, G., Zink, C. F., Montague, P. R., & Berns, G. S. (2002). Activity in human ventral striatum locked to errors of reward prediction. *Nature Neuroscience*, *5*(2), 97-98.

Parent, A., & Hazrati, L. N. (1995). Functional anatomy of the basal ganglia. I. The cortico-basal ganglia-thalamo-cortical loop. *Brain Research Reviews*, *20*(1), 91-127.

Parkinson, J. A., Willoughby, P. J., Robbins, T. W., & Everitt, B. J. (2000). Disconnection of the anterior cingulate cortex and nucleus accumbens core impairs pavlovian approach behavior: Further evidence for limbic cortical–ventral striatopallidal systems. *Behavioral Neuroscience*, *114*(1), 42-63.

Parr, R. E. (1998). *Hierarchical control and learning for markov decision processes* (Unpublished doctoral dissertation). University of California.

Parr, R. E., & Russell, S. (1998). Reinforcement learning with hierarchies of machines. In M. S. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in neural information processing systems* (pp. 1043–1049). Cambridge, MA: MIT 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*(6), 532-552.

Pearce, J. M., Kaye, H., & Hall, G. (1982). Predictive accuracy and stimulus associability: development of a model for pavlovian learning. In M. L. Commons, R. J. Herrnstein, & W. A. R (Eds.), *Quantitative*

analysis of behavior: Acquisition (Vol. 3, pp. 241–256). Cambridge, MA: Ballinger Publishing Company.

Phan, K. L., Wager, T. D., Taylor, S. F., & I, L. (2004). Functional neuroimaging studies of human emotions. *CNS Spectrums*, *9*, 258-266.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, *10*(2), 59-63.

Ponsen, M., Taylor, M. E., & Tuyls, K. (2010). Abstraction and generalization in reinforcement learning: A summary and framework. In M. E. Taylor & K. Tuyls (Eds.), *Adaptive and learning agents* (pp. 1–32). Berlin: Springer.

Postle, B. R. (2006). Working memory as an emergent property of the mind and brain. *Neuroscience*, *139*(1), 23-28.

Precup, D. (2000). *Temporal abstraction in reinforcement learning* (Unpublished doctoral dissertation). University of Massachusetts.

Raab, M., & Johnson, J. G. (2007). Expertise-based differences in search and option-generation strategies. *Journal of Experimental Psychology: Applied*, *13*(3), 158-170.

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences*, *98*(2), 676–682.

Ravel, S., Sardo, P., Legallet, E., & Apicella, P. (2006). Influence of spatial information on responses of tonically active neurons in the monkey striatum. *Journal of Neurophysiology*, *95*(5), 2975-2986.

Reason, J. T. (1979). Actions not as planned: The price of automatization. In G. Underwood & R. Stevens (Eds.), *Aspects of consciousness* (p. 67-89). London: Academic Press.

Reed, E. S., Montgomery, M., Schwartz, M., Palmer, C., & Pittenger, J. B. (1992). Visually Based Descriptions of an Everyday Action. *Ecological Psychology*, *29*(7), 690–705.

Reed, P., Mitchell, C., & Nokes, T. (1996). Intrinsic reinforcing properties of putatively neutral stimuli in an instrumental two-lever discrimination task. *Animal Learning & Behavior*, *24*(1), 38–45.

Reynolds, J. R., & Mozer, M. C. (2009). Temporal dynamics of cognitive control. In *Advances in neural information processing systems.* Cambridge, MA: MIT Press.

Reynolds, J. R., O'Reilly, R. C., Cohen, J. D., & Braver, T. S. (2012). The function and organization of lateral prefrontal cortex: A test of competing hypotheses. *Plos One*, *7*(2), e30284.

Reynolds, J. R., Zacks, J. M., & Braver, T. S. (2007). A computational model of event segmentation from perceptual prediction. *Cognitive Science*, *31*(4), 613–643.

Ribas-Fernandes, J. J. F., Niv, Y., & Botvinick, M. M. (2011). Neural correlates of Hierarchical Reinforcement Learning. In R. B. Mars, J. Sallet, M. F. S. Rushworth, & N. Yeung (Eds.), *Neural basis of motivational and cognitive control* (p. 285-309). Cambridge, MA: MIT Press.

Ribas-Fernandes, J. J. F., 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*(2), 370–379.

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., & Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science*, *306*(5695), 443–447.

Roesch, M. R., Esber, G. R., Li, J., Daw, N. D., & Schoenbaum, G. (2012). Surprise! Neural correlates of Pearce-Hall and Rescorla-Wagner coexist within the brain. *European Journal of Neuroscience*, *35*(7), 1190–1200.

Roesch, M. R., Taylor, A. R., & Schoenbaum, G. (2006). Encoding of time-discounted rewards in orbitofrontal cortex is independent of value. *Neuron*, *51*(4), 509-520.

Rohanimanesh, K., & Mahadevan, S. (2001). Decision-theoretic planning with concurrent temporally extended actions. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence* (p. 472-279). Morgan Kauffmann.

Rosenbaum, D. A., Kenny, S. B., & Derr, M. A. (1983). Hierarchical control of rapid movement sequences. *Journal of Experimental Psychology: Human Perception and Performance*, *9*(1), 86-102.

Rougier, N. P., Noell, D. C., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (2005). Prefrontal cortex and flexible cognitive control: rules without symbols. *Proceedings of the National Academy of Sciences*, *102*(20), 7338-7343.

Rushworth, M. F. S., Noonan, M. P., Boorman, E. D., Walton, M. E., &

Behrens, T. E. (2011). Frontal cortex and reward-guided learning and decision-making. *Neuron*, *70*(6), 1054–1069.

Rushworth, M. F. S., Walton, M. E., Kennerley, S. W., & Bannerman, D. M. (2004). Action sets and decisions in the medial frontal cortex. *Trends in Cognitive Sciences*, *8*(9), 410-417.

Rutledge, R. B., Dean, M., Caplin, A., & Glimcher, P. W. (2010). Testing the reward prediction error hypothesis with an axiomatic model. *Journal of Neuroscience*, *30*(40), 13525-13536.

Saffran, J. R., & Wilson, D. P. (2003). From syllables to syntax: multilevel statistical learning by 12-month-old infants. *Infancy*, *4*(2), 273–284.

Saga, Y., Iba, M., Tanji, J., & Hoshi, E. (2011). Development of multidimensional representations of task phases in the lateral prefrontal cortex. *Journal of Neuroscience*, *31*(29), 10648–10665.

Sakai, K., & Passingham, R. E. (2006). Prefrontal set activity predicts rule-specific neural processing during subsequent cognitive performance. *Journal of Neuroscience*, *26*(4), 1211-1218.

Salas, R., & Montague, P. R. (2010). BOLD Responses to Negative Reward Prediction Errors in Human Habenula. *Frontiers in Human Neuroscience*, *4*.

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature Neuroscience*, *16*, 486-492.

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

Schultz, W., Tremblay, K. L., & Hollerman, J. R. (2000). Reward processing in primate orbitofrontal cortex and basal ganglia. *Cerebral Cortex*, *10*(3), 272-283.

Schwartz, M. F., Montgomery, M. W., Fitzpatrick-DeSalme, E. J., Ochipa, C., Coslett, H. B., & Mayer, N. H. (1995). Analysis of a disorder of everyday action. *Cognitive Neuropsychology*, *12*(8), 863–892.

Schwartz, M. F., Reed, E. S., Montgomery, M., Palmer, C., & Mayer, N. H. (1991). The Quantitative Description of Action Disorganisation after Brain Damage: A Case Study. *Cognitive Neuropsychology*, *8*(5), 381–414.

Seo, H., & Lee, D. (2007). Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game. *Journal of*

Neuroscience, *27*(31), 8366-8377.

Seymour, B., O'Doherty, J. P., Dayan, P., Koltzenburg, M., Jones, A. K., Dolan, R. J., ... Frackowiak, R. S. (2004). Temporal difference models describe higher-order learning in humans. *Nature*, *429*(6992), 664–667.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217-240.

Shima, K., Isoda, M., Mushiake, H., & Tanji, J. (2007). Categorization of behavioural sequences in the prefrontal cortex. *Nature*, *445*(7125), 315–318.

Shima, K., Mushiake, H., Saito, N., & Tanji, J. (1996). Role for cells in the presupplementary motor area in updating motor plans. *Proceedings of the National Academy of Sciences*, *93*(16), 8694-8698.

Shima, K., & Tanji, J. (2000). Neuronal activity in the supplementary and presupplementary motor areas for temporal organization of multiple movements. *Journal of Neurophysiology*, *84*(4), 2148–2160.

Shimamura, A. P. (2000). The role of the prefrontal cortex in dynamic filtering. *Psychobiology*, *28*(2), 207-218.

Şimşek, Ö., & Barto, A. G. (2004). Using relative novelty to identify useful temporal abstractions in reinforcement learning. In *Proceedings of the twenty-first international conference on machine learning* (p. 95). New York, NY.

Şimşek, Ö., Wolfe, A. P., & Barto, A. G. (2005). Identifying useful subgoals in reinforcement learning by local graph partitioning. In *Proceedings of the 22nd international conference on machine learning* (pp. 816–823).

Singh, S., Barto, A. G., & Chentanez, N. (2005). Intrinsically motivated reinforcement learning. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems* (p. 1281-1288). Cambridge, MA: MIT Press.

Sirigu, A., Zalla, T., Pillon, B., Dubois, B., Grafman, J., & Agid, Y. (1995). Selective impairments in managerial knowledge following pre-frontal cortex damage. *Cortex*, *31*(2), 301-316.

Solway, A., Diuk, C., Cordova, N., Yee, D., Barto, A. G., Niv, Y., & Botvinick, M. M. (submitted). *Optimal behavioral hierarchy.*

Spence, K. W., & Grice, G. R. (1942). The role of final and sub-goals in distance discrimination by the white rat. *Journal of Comparative Psychology*, *34*(2), 179–184.

Suri, R. E., Bargas, J., & Arbib, M. A. (2001). Modeling functions of striatal dopamine modulation in learning and planning. *Neuroscience*, *103*(1), 65–86.

Sutton, R., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, *112*(1-2), 181-211.

Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of pavlovian reinforcement. In M. Moore & J. Gabriel (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks* (p. 497-537). Cambridge, MA: MIT Press.

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

Szabo, J. (1979). Strionigral and nigrostriatal connections. anatomical studies. *Applied Neurophysiology*, *42*, 9–12.

Szepesvari, C. (2010). Algorithms for reinforcement learning. In R. J. Brachman & T. G. Dietterich (Eds.), *Synthesis lectures on artificial intelligence and machine learning.* doi: 10.2200/S00268ED1V01Y201005AIM009: Morgan and Claypool.

Talairach, J., & Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain.* New York, NY: Thieme.

Talmi, D., Atkinson, R., & El-Deredy, W. (2013). The feedback-related negativity signals salience prediction errors, not reward prediction errors. *Journal of Neuroscience*, *33*(19), 8264–8269.

Taylor, M. E., & Stone, P. (2009). Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, *10*, 1633–1685.

Terrace, H. S. (1993). The phylogeny and ontogeny of serial memory: list learning by pigeons and monkeys. *Psychological Science*, *4*(3), 162-169.

Thelen, E. (1981). Rhythmical behavior in infancy: An ethological perspective. *Developmental Psychology*, *17*(3), 237-257.

Thrun, S., & Schwartz, A. (1995). Finding structure in reinforcement learning. In *Advances in neural information processing systems 7.*

Cambridge, MA: MIT Press.

Tremblay, P.-L., Bedard, M.-A., Langlois, D., Blanchet, P. J., Lemay, M., & Parent, M. (2010). Movement chunking during sequence learning is a dopamine-dependant process: a study conducted in parkinson's disease. *Experimental Brain Research*, *205*(3), 375–385.

Tricomi, E., Balleine, B. W., & ODoherty, J. P. (2009). A specific role for posterior dorsolateral striatum in human habit learning. *European Journal of Neuroscience*, *29*(11), 2225–2232.

Turk-Browne, N. B., & Scholl, B. J. (2010). Statistical learning. In B. Goldstein (Ed.), *Encyclopedia of perception* (p. 935-938). Thousand Oaks, CA: Sage Publications.

Uithol, S., van Rooij, I., Bekkering, H., & Haselager, P. (2012). Hierarchies in action and motor control. *Journal of Cognitive Neuroscience*, *24*(5), 1077–1086.

Ullsperger, M., & von Cramon, D. Y. (2003). Error monitoring using external feedback: specific roles of the habenular complex, the reward system, and the cingulate motor area revealed by functional magnetic resonance imaging. *Journal of Neuroscience*, *23*(10), 4308–4314.

Van Dijk, S. G., Polani, D., & Nehaniv, C. L. (2011). Hierarchical behaviours: getting the most bang for your bit. In *Advances in artificial life. Darwin Meets von Neumann* (pp. 342–349). Springer.

van Hasselt, H. (2012). Reinforcement learning in continuous state and action spaces. In M. A. Wiering & M. van Otterlo (Eds.), *Reinforcement learning: State of the art* (Vol. 12, p. 207-251). Berlins: Springer.

van Veen, V., Holroyd, C. B., Cohen, J. B., Stenger, V. A., & Carter, C. S. (2004). Errors without conflict: Implications for performance monitoring theories of anterior cingulate cortex. *Brain and Cognition*, *56*(2), 267–276.

Viard, A., Doeller, C. F., Hartley, T., Bird, C. M., & Burgess, N. (2011). Anterior hippocampus and goal-directed spatial decision making. *Journal of Neuroscience*, *31*(12), 4613–4621.

Wallis, J. D., Anderson, K. C., & Miller, E. K. (2001). Single neurons in prefrontal cortex encode abstract rules. *Nature*, *411*(6840), 953-956.

Walton, M. E., Devlin, J. T., & Rushworth, M. F. (2004). Interactions between decision making and performance monitoring within prefrontal cortex. *Nature Neuroscience*, *7*(11), 1259–1265.

Walton, M. E., Groves, J., Jennings, K. A., Croxson, P. L., Sharp, T., Rushworth, M. F., & Bannerman, D. M. (2009). Comparing the role of the anterior cingulate cortex and 6-hydroxydopamine nucleus accumbens lesions on operant effort-based decision making. *European Journal of Neuroscience*, *29*(8), 1678–1691.

Ward, T. B. (2007). Creative cognition as a window on creativity. *Methods*, *42*(1), 28–37.

Weiner, M. J., Hallett, M., & Funkenstein, H. H. (1983). Adaptation to lateral displacement of vision in patients with lesions of the central nervous system. *Neurology*, *33*(6), 766–766.

White, I. M., & Wise, S. P. (1999). Rule-dependent neuronal activity in the prefrontal cortex. *Experimental Brain Research*, *126*(3), 315-335.

Whiten, A., Flynn, E., Brown, K., & Lee, T. (2006). Imitation of hierarchical action structure by young children. *Developmental Science*, *9*(6), 574–582.

Wickens, J., Kotter, R., & Houk, J. C. (1995). Cellular models of reinforcement. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (p. 187-214). Cambridge, MA: MIT Press.

Wiering, M., & Schmidhuber, J. (1998). HQ-learning. *Adaptive Behavior*, *6*(2), 219-246.

Wingate, D., Diuk, C., O'Donnell, T., Tenenbaum, J. B., & Gershman, S. J. (2013). *Compositional policy priors* (Tech. Rep. No. 2013-007). MIT CSAIL.

Wise, R. A. (2002). Brain reward circuitry: insights from unsensed incentives. *Neuron*, *36*(2), 229-240.

Witten, I. B., Steinberg, E. E., Lee, S. Y., Davidson, T. J., Zalocusky, K. A., Brodsky, M., ... others (2011). Recombinase-driver rat lines: tools, techniques, and optogenetic application to dopamine-mediated reinforcement. *Neuron*, *72*(5), 721–733.

Wood, J. N., & Grafman, J. (2003). Human prefrontal cortex: processing and representational perspectives. *Nature Reviews Neuroscience*, *4*(2), 139–147.

Yacubian, J., Gläscher, J., Schroeder, K., Sommer, T., Braus, D. F., & Büchel, C. (2006). Dissociable systems for gain- and loss-related value predictions and errors of prediction in the human brain. *Journal of*

Neuroscience, *26*(37), 9530–9537.

Yamaguchi, S., Tsuchiya, H., & Kobayashi, S. (1995). Electrophysiologic correlates of visuo-spatial attention shift. *Electroencephalography and Clinical Neurophysiology*, *94*(6), 450–461.

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychological Review*, *111*(4), 931–959.

Yeung, N., Holroyd, C. B., & Cohen, J. D. (2005). ERP correlates of feedback and reward processing in the presence and absence of response choice. *Cerebral Cortex*, *15*(5), 535-544.

Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, *7*(6), 464–476.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience*, *19*(1), 181-189.

Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., ... Raichle, M. E. (2001). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience*, *4*(6), 651-655.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: a mind-brain perspective. *Psychological Bulletin*, *133*(2), 273-293.

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin*, *127*(1), 3-21.

Zalla, T., Pradat-Diehl, P., & Sirigu, A. (2003). Perception of action boundaries in patients with frontal lobe damage. *Neuropsychologia*, *41*(12), 1619-1627.

ITQB-UNL | Av. da República, 2780-157 Oeiras, Portugal
Tel (+351) 214 469 100
Fax (+351) 214 411 277

www.itqb.unl.pt