Université Pierre et Marie Curie (UPMC)
Institute of Intelligent Systems and Robotics

Habilitation to Direct Researches (HDR)

speciality « Biology »

by

Mehdi Khamassi

Coordination of parallel learning processes in animals and robots

HDR defended on 6th May 2014 in front of the jury composed of:

Dr.	Frédéric ALEXANDRE	INRIA	(Reviewer)
Dr.	Raja CHATILA	CNRS – UPMC	(Examiner)
Dr.	Boris S. GUTKIN	CNRS – ENS	(Reviewer)
Dr.	Mathias PESSIGLIONE	INSERM	(Examiner)
Pr.	Tony J. PRESCOTT	Univ. Sheffield	(Reviewer)
Dr.	Emmanuel PROCYK	CNRS – INSERM	(Examiner)

Acknowledgments

I would like to thank the members of the jury for accepting to evaluate this work, for having taken the time to read this manuscript, and for coming to ISIR in Paris for the defense.

I would also like to thank my collaborators, without whom all this work would not have been possible.

All my gratitude to my family and friends for their support throughout these slightly more than 10 years of research, since the beginning of my PhD at the end of 2003.

Spéciale casse-dédi to Alexandra and Thalia Leila.

Trento, Italy, 3rd March 2014.

Table of contents

List of figures

Introduction

1

Contents

This HDR manuscript presents research work at the interface between Computational Neuroscience and Cognitive Robotics. The main scientific issue at stake is to understand how animals and robots can display behavioral adaptation capabilities in their partially unknown and changing environment. The objective is two-fold : on the one hand, contributing to better understanding behavioral and neural correlates of learning processes ; on the other hand, taking inspiration from biology to design autonomous robots able to learn from their own observations and errors. This work is built on previous evidence that the mammalian brain combines different memory systems which enable parallel learning processes for efficient behavioral adaptation. Within the instrumental conditioning paradigm, this is reflected by initial goal-directed learning observed in animals which seem to build and use an internal model of their environment, followed by the progressive expression of habits that have been slowly learned in parallel. In computational terms, this can be formalized as a progressive shift from model-based (MB) to model-free (MF) reinforcement learning (RL). In the navigation paradigm, this is reflected through the alternation between different navigation strategies, which can also be categorized into MB and MF RL processes. The manuscript presents work

performed with collaborators – among whom supervised PhD students – to contribute to : 1) Proposing computational solutions for the coordination of parallel learning processes to explain animal behavior during conditioning and navigation paradigms ; 2) Using learning models to analyze behavioral dynamics and neural activities recorded in animals during behavioral adaptation ; 3) Implementing neuro-inspired learning models in robots to make them work in the real world. An emphasis is put on the add-ons and gains produced by these exchanges between disciplines and approaches. In particular, the manuscript highlights (i) how computational models can help better formalize and quantify information processes that may underlie animal behavior and brain activity ; (ii) how neuro-inspired models can constitute a complementary and fruitful approach to classical Robotics work ; (iii) in return how Robotics implementations can help improve neurocomputational models by testing their robustness in the real world, by discovering new properties of these models in such conditions, and by raising new questions and hypotheses concerning the necessary coordination between learning processes to properly work in a physical body. Finally, a discussion of possible future directions of investigations is proposed in order to plan a research program for the forthcoming years.

1.1 Scientific context

1.1.1 Different machine learning algorithms

In the field of Machine Learning, learning algorithms can be roughly categorized into three main groups, depending on the feedback that the learner receives :

- **Supervised Learning**, where the feedback tells exactly which target output the learner should have generated in response to the input. This type of learning is typically used when a neural network model learns to recognize handwritten characters and is corrected when its guess is different from the known true character at a given trial.
- **Reinforcement Learning**, where the feedback does not tell what the target output was but just says whether the output generated by the learner is good, bad or neutral. This is typically concerned with situations where an agent has to learn how to act in an environment in order to maximise some notion of reward.
- **Unsupervised Learning**, where no feedback is received by the learner. In this case, the learner typically has to learn the data structure, *e.g.* learning that some elements are always associated together, or learning the regular temporal contiguity between a couple of events.

In an influential Computational Neuroscience paper in 2000, Kenji Doya made the proposition that different brain regions, namely the Cerebellum, the Basal Ganglia, and the Cortex, are each mainly involved in one among these three different types of learning processes (Fig. 1.1). Although quite schematic, this view is still highly relevant today. Recent computational models of cerebellar function still emphasize the predominant role of supervised learning in this structure (Kawato et al. 2011). Reinforcement learning continues to play a central role in basal ganglia

(BG) models (Dayan and Niv 2008, Maia and Frank 2011, Keramati and Gutkin 2013). And the associative nature of cortical networks makes unsupervised learning a key process in many models of the cortex's role in decision-making (Hasselmo 2005, Martinet et al. 2011) or in other cognitive functions (Fix et al. 2007).

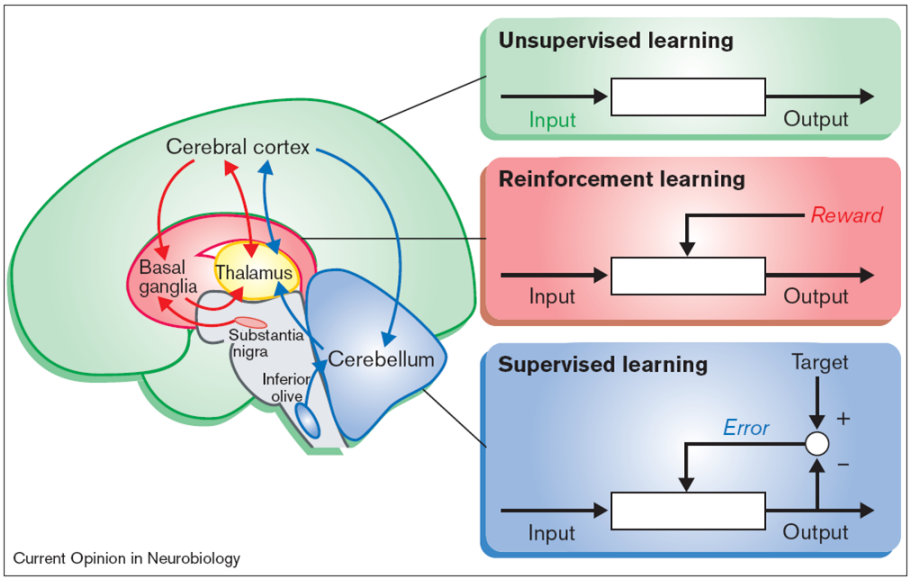


FIGURE 1.1 – *Proposition of a decomposition of neural structures involved in different types of learning processes (Doya (2000a) with permissions)*

The work presented in this manuscript mostly focuses on reinforcement learning (RL) and unsupervised learning (UL) processes, and on their corresponding neural substrates in the basal ganglia and cortex. The motor control part is not addressed, which means that the computational models presented are most of the time simplified by manipulating abstract actions without wondering how sequences of muscle activations are learned and organized for the execution of these actions. Incorporating the motor control part in the models would also require a proper coordination of learning processes by itself (*i.e.* between supervised learning and other learning processes). This has been left aside for the moment. Nevertheless a first step in this direction is sketched in the long-term research project presented at the end of this manuscript (Section 5.2).

The present work addresses the question of how to coordinate different RL processes together and how to coordinate RL and UL processes together in order to produce efficient and biologically realistic behavioral adaptation abilities. In the case of animal and robot learning, UL processes are important to learn an internal model which incorporates information about the structure of the environment or of the task. For instance a cognitive map containing topological links between diferent locations within the environment and enabling to plan the shortest path towards a goal position. Or a graph of transitions between states of an instrumental conditioning task, enabling to plan a sequence of decisions until a desired state, and permitting to avoid sequences of actions that lead to a long-term undesired state (*e.g.* a devalued outcome). Throughout the manuscript, a distinction will often be mentioned between *model-based reinforcement learning (MBRL)* – when the learning process includes the build-up and the use

of such an internal model – and *model-free reinforcement learning (MFRL)* – when the learning process occurs without access to such a model. These two subclasses of RL processes will be more precisely and formally defined in the forthcoming sections (*e.g.* see equations and schemes in the paper Khamassi and Humphries (2012) presented in Section 2.1.1). The important thing to remember at this stage is that this distinction will be determinant to characterize different types of learning behaviors and their underlying neural substrates.

1.1.2 Reinforcement learning and animal behavioral adaptation

Animals' ability to learn from their own experience and errors, in particular in the context of sparse reward and punishment signals, crucially relies on reinforcement learning processes. The most central theory currently considers that such learning relies on : 1) the competition between actions, resulting in action selection as a function of the actions' relative probabilities ; 2) the anticipation of the value of rewards and punishments that could follow the execution of the action ; 3) the computation of a *reward prediction error* comparing what was expected with what is actually obtained ; 4) the use of such a reward prediction error as a feedback (*i.e.* positive, negative or null reinforcement signal) to update either the probability of the performed action or the predictive value associated to the action and to the stimuli present in this context (Sutton and Barto 1998).

This formalism can be seen as an extension of the Rescorla-Wagner model (Rescorla and Wagner 1972) in which learning requires prediction errors to explain various properties of associative learning during animals classical conditioning. Prediction errors can indeed explain the *blocking* phenomenon – when a stimulus B cannot be associated with a reward if it is presented together with a stimulus A which is already fully predictive of the reward –, and cases of *overexpectation* – when the concomittant presentation of two reward predictive stimuli influences behavior as if they were adding up, to form a stronger prediction.

A particular subgroup of RL algorithms implementing what is called Temporal-Difference (TD) learning extend the Rescorla-Wagner model in that prediction error signals contain three terms rather than two. The Rescorla-Wagner indeed compares past expectation with present outcome (*e.g.* reward). The TD learning rule adds to this comparison a term representing future expectations of reward (see equations in the paper Khamassi and Humphries (2012) presented in Section 2.1.1). As a consequence, a reinforcement signal can be computed even before the reward is attained by comparing temporally consecutive expectations of reward – hence the term *Temporal-Difference* : *e.g.* when an action leads to a situation or state where reward expectations are higher than previous ones, this action should be reinforced.

1.1.3 Applications to Neuroscience

Since nearly twenty years, this theory has provided Neuroscientists with formal tools which contributed to important breakthroughs in the understanding of neural correlates of learning. Reinforcement Learning

models turned out to be able to explain a wide range of adaptive behaviors experimentally observed both in humans (Rushworth and Behrens 2008, Frank et al. 2009, Balleine and O'Doherty 2010, Collins and Koechlin 2012) and in non-human animals (*e.g.* Yin and Knowlton (2006)). This formalism also enabled to explain a variety of neural correlates of learning (Schultz et al. 1997, Yin et al. 2008). The most striking example and probably the most central in the field is the observation that phasic responses of dopaminergic neurons follow the profile of reward prediction errors as they are formalized by the RL theory : an increase in activity when the outcome of action is better than expected ; a decrease in activity when it is worse than expected ; an absence of response when it meets the expectations (Schultz et al. 1997, Bayer and Glimcher 2005, Morris et al. 2006, Roesch et al. 2007, Matsumoto and Hikosaka 2009).

The accumulation of neurophysiological results corroborated by this computational theory has also enabled to establish that the learning of reward values and action values depends on plasticity in projections from the cortex to the basal ganglia (BG ; in particular to the striatum), and that these adjusments depend on dopaminergic signals sent from the substantia nigra pars compacta (SNc) and the ventral tegmental area (VTA) (Houk et al. 1995, Schultz et al. 1997, Doya 2000a, Reynolds et al. 2001, O'Doherty et al. 2004, Samejima et al. 2005, Faure et al. 2005, Pessiglione et al. 2006, Shen et al. 2008, Humphries and Prescott 2010, van der Meer and Redish 2011). Numerous computational models of the basal ganglia (BG) were derived from these experimental results (Houk et al. 1995, Schultz et al. 1997, Doya 2000a, Joel et al. 2002, Baldassarre 2002, Frank 2005), and were built on the central assumption that the BG play a critical role in action selection (Redgrave et al. 1999, Gurney et al. 2001).

My PhD work (Khamassi 2007) contributed in showing that ventral striatal single-unit activity in behaving rats is coherent with the RL theory (Khamassi et al. 2008) and in constraining BG RL models to make them physiologically and anatomically plausible as well as efficient in realistic continuous simulations of laboratory tasks (Khamassi et al. 2005; 2006). I have also been recently collaborating with Mark D. Humphries and Kevin Gurney to propose a more recent RL model of the BG which incorporates a role for tonic dopamine in the regulation the exploration-exploitation trade-off for action selection (Humphries et al. 2012).

The application of the RL theory to Neuroscience also favored the development of a method for *model-based* analysis of experimental data (Daw et al. 2006, Corrado and Doya 2007, Brovelli et al. 2008, Ito and Doya 2009, Palminteri et al. 2009, Daw 2011, Collins and Koechlin 2012). In this approach, a computational model is parametrized in order to fit subjects' observed behavior during the task with a Bayesian maximum likelihood criterion. Then hidden model variables are used as regressors of the recorded neural activity to test the assumption that this activity reflects computations similar to those performed by the model to solve the task.

Chapter 3 in this HDR manuscript presents two studies using such a method for model-based analyses of neurophysiological data : one started during my postdoctoral training with Emmanuel Procyk and Peter F. Dominey (Khamassi et al. 2014) ; the other done by Jean Bellot, a PhD student that I co-supervise with Benoît Girard, in collaboration with Oli-

vier Sigaud, Geoffrey Schoenbaum and Matthew R. Roesch (Bellot et al. in preparation).

Nevertheless, the results obtained through the application of the RL theory to Neuroscience remain fragmentary and incomplete for several reasons. First, the laboratory tasks employed are most of the time very simple, involving only a few different stimuli and actions. Second, these tasks most of the time involve single-step decisions performed at each trial, while the RL theory has been designed to deal with multiple steps of decision-making and reward predictions. Finally, these studies – including some of ours – most of the time use a single computational model to explain behavior (but see Gläscher et al. (2010), Daw et al. (2011)), while there is more and more evidence that subjects' behavior during decision-making tasks involve the coordination of multiple learning systems (Daw et al. 2005).

1.1.4 Coordination of multiple learning modules

Rodents put in an instrumental conditioning paradigm – where they have to learn to press a lever in response to a cue in order to get reward – initially display flexible learning behavior and progressively develop habits that are long and difficult to break (Dickinson 1985). After moderate training, changes in contingencies of the task or devaluation of the reward – for instance by pairing it with illness – result in relatively fast behavioral adaptation : the animal quickly stops pressing the lever. In contrast, after extensing training animals persist in pressing the lever in this context no manner if the task contingencies have changed or if the reward has been devalued.

Such behavioral flexibility and automaticity are associated with two separate learning systems : the goal-directed and habit learning systems (Balleine and O'Doherty 2010). These systems are mediated by separate cortico-striatal networks, namely the associative and sensorimotor fronto-striatal loops, respectively. The associative loop includes the lateral and medial prefrontal cortices and the dorsomedial striatum of the basal ganglia, whereas the sensorimotor circuit includes sensorimotor and premotor areas that project to the dorsolateral striatum (Yin and Knowlton 2006, Graybiel 2008, Ashby et al. 2010). Such dual-system hypothesis conforms the notion that frontal cortex activity is organised according to a rostro-caudal gradient based on the abstractness of action representations (Koechlin et al. 2003, Badre and D'Esposito 2009), assigning goal-directed actions to anterior portions of the frontal lobe and stimulus-response habits to sensorimotor areas.

At the theoretical level, the reinforcement learning theory (Sutton and Barto 1998) is providing a coherent mathematical framework to formalize goal-directed and habit learning computations (Dayan and Balleine 2002, Daw et al. 2005, Ito and Doya 2011). In particular, Daw and colleagues (Daw et al. 2005) proposed that a dual learning system involving model-based and model-free reinforcement learning algorithms, the former employing "effortful" computations in a model of the world (*i.e.*, goal-directed learning), the latter producing reactive behaviors based on stimulus-response associations (*i.e.*, habit learning). Model-based RL me-

chanisms are good models of goal-directed behavior because they involve a model of long-term consequences of actions which enable to plan ahead and to avoid sequences of actions that lead to non-desired goals (*e.g.* devalued goals). They thus produce flexible behavior in response to changes in the environment because a single exposure to changes in the goal permits a change in subsequent decision-making. However, such planning with the model-based system before acting produces slow reaction times and is computational costly. In contrast, model-free RL mechanisms explain habit learning in that Temporal-Difference learning algorithms slowly propagate value information from the end of the action sequence (*i.e.* where the agent gets reward) to the beginning of the sequence. Hence the slowness to acquire a habit and the even longer time required to break a habit – because the negative value associated to a devalued reward will first need to decrease the positive values associated to each elements within the action sequence before properly being able to propagate negative values to the whole sequence. However, making a decision with the model-free system is much quicker and thus produces slow reaction times because one "just" needs to compare a small set of cached values associated to the actions in competition [1].

Interestingly, a recent paper suggests that habit learning may be better modeled by a chunking mechanism – automatizing the selection of frequently repeated sequences of actions – rather than by a classical TD-learning algorithm (Dezfouli and Balleine 2012). Although this is something we started to investigate in Robotics by comparing the behavioral properties produced by both systems in realistic continuous situations, this work is preliminary and will not be presented in this manuscript. Nevertheless, this shows that the debate concerning the precise nature and computational mechanisms underlying habit learning is still vivid. There is also an important debate concerning the possible mechanisms underlying the coordination of reinforcement learning systems. Daw et al. (2005) proposed an uncertainty-based mechanism for this coordination : the system that computes reward values with the lowest uncertainty takes over behavior. However, this method requires costly calculations of the uncertainty in the two systems – while the capacity to learn habits may have emerged through evolution to enable computation saving by avoiding to systematically use the goal-directed system (Killcross and Coutureau 2003). Moreover the complexity of uncertainty computation within the model-based system makes it exponentially explode with the number of states. In the simple task with six states simulated by Daw et al. (2005), this is not a problem. But in more realistic situations with a large number of states, this computation becomes problematic. To cope with this issue, the model of Keramati et al. (2011) proposes to only compute the less expensive uncertainty of the model-free system, and to avoid using the model-based system when this uncertainty is low. On the one hand this model enables to save computation time and to explain a substantial set of experimental data. On the other hand, this model relies on the simplied assumption that the model-based system is always more reliable

1. It is worthy of note that Robotic experiments with continuous action spaces show that in such a case the comparison between action values is much more difficult and slower (Peters and Schaal 2006, van Hasselt and Wiering 2007).

and less uncertain than the model-free system, which is not always the case in more realistic and embodied situations such as the robotic experiments presented in Chapter 4 of this manuscript. Depending on noise, uncertainty and characteristics of the environment and of the robot's perceptual equipment, it turns out that most of the time the model-based system is more efficient than the model-free one in some parts of the environement and vice-versa in other parts. It appears thus promising to use system coordination criteria without too many priors and with rather the ability to automously detect which system is the most appropriate in each circumstance.

Several contributions on the modelling of the coordination of learning systems are presented in this HDR manuscript. In Chapter 2, two papers are presented showing that the model-based / model-free dichotomy is also relevant (i) to categorize different navigation strategies observed in rodents and the activity of their underlying neural substrates (work done in collaboration with Mark D. Humphries, Khamassi and Humphries (2012)), (ii) and to explain inter-individual differences in rats' behavior during Pavlovian conditioning – differences in behaviors called *goal-tracking* versus *sign-tracking* – as well as differences in dopamine activity observed in these rats (work done by Florian Lesaint, a PhD student that I co-supervise with Olivier Sigaud, in collaboration with Shelly B. Flagel and Terry E. Robinson, Lesaint et al. (2014)). Besides, Chapter 4 shows robotic implementations of a model-based / model-free computational model (Dollé et al. 2008; 2010; submitted) for robot navigation (work done by Ken Caluwaerts, a Master student that I co-supervised with Agnès Guillot and Christophe Grand, Caluwaerts et al. (2012b)). This model has the advantage of having a memory of which learning system was the most efficient in each subpart of the environment, which can produce faster recovery of the most reliable system at each moment, and which property was absent from previous computational models of the coordination of model-based / model-free systems.

1.1.5 Meta-learning and cognitive control

Interestingly, the question of how to efficiently coordinate multiple learning systems is the subject of investigations within the Machine Learning literature, within a subfield called *meta-learning* (Schmidhuber et al. 1997, Doya 2002, Giraud-Carrier et al. 2004). It is a concept originally developed within the domain of Cognitive Psychology which means learning to learn. In Machine Learning the term refers to applications of learning algorithms to meta-data in order to find out what mechanisms and principles can reveal flexible and general enough to solve different kinds of problems.

A major issue in Machine Learning is indeed concerned with algorithm parametrization, often performed specifically for the task to solve, thus enabling little generalization to different tasks (Lavesson and Davidsson 2006). Within the context of Markov Decision Problems (MDP) – where an agent has to learn a behavioral policy in order to maximize a given reward function –, the parameters tuned enable the tested reinforcement learning algorithms to solve a particular condition do not permit rapid behavioral

adaptation once the task conditions are changed (Sutton and Barto 1998). There is thus a need for meta-heuristics enabling to dynamically optimise the parameters of the algorithms, and to permit simulated agents to take appropriate decisions and learn in unprepared, changing environments. Some existing methods such as evolutionary algorithms enable off-line optimization in the sense that the latter does not occur during the lifetime of the agent (Doncieux et al. 2011). However, if the goal is to enable on-line incremental adaptation, meta-learning solutions can be appropriate.

From a mechanistic point of view, the interesting thing is that meta-learning methods not only have been proposed (i) to select the appropriate model for action selection in the case of the coordination of multiple learning systems (Brazdil 1998), but also (ii) to dynamically regulate parameters of learning (*e.g.* update rate, temporal scale and exploration parameters, Auer et al. (2002), Ishii et al. (2002)). So the computational mechanisms underlying animals' behavioral flexibility and adaptivity could be both investigated in terms of learning module coordination and dynamic regulation of learning parameters.

During my post-doctoral work in collaboration with Emmanuel Procyk and Peter F. Dominey, we have drawn a parallel between meta-learning principles proposed in Machine Learning and cognitive control processes described in Neuroscience – *i.e.* how to regulate the appropriate level of control to solve a given task – (Khamassi et al. 2011b; 2013). Recent reviews of neurobiological data have indeed highlighted cognitive control mechanisms for the high-level coordination of executive systems in the primate prefrontal cortex (Miller and Cohen 2001, Koechlin and Summerfield 2007, Samejima and Doya 2007). The cognitive control loop theory describes the modulation of the control level enabling shifting from routine behaviors in a known context requiring little attention and concentration, to more flexible behaviors involving rapid and active control. The level of control is based on a monitoring process of variations of the environment and of the agent's own performance. It also implies learning the association between particular tasksets and the contexts in which they are relevant.

During my PhD work in collaboration with Sidney I. Wiener, Francesco P. Battaglia, Adrien Peyrache, Karim Benchenane, Yves Gioanni and Patrick L. Tierney, I contributed to electrophysiological recordings in the Hippocampus-Prefrontal Cortex network in rats performing a decision-making task in a Y-maze, with regularly changing task rules. We found that the activity of cell assemblies within this network reflected a learning process of which task-rule (*i.e.* taskset) is currently appropriate to solve the task, these activities being related to an increase in the coherence between Hippocampus and Prefrontal Cortex activities at decision time when the current task-rule has been discovered by the animal, and being replayed during sleep, putatively enabling a consolidation of this knowledge (Battaglia et al. 2008, Peyrache et al. 2009; 2010a;b, Benchenane et al. 2010). These data provide us with some clues about the possible mechanisms underlying the prefrontal cortex's involvment in cognitive control.

Thus a fruitful approach can consist in both (i) investigating whether some meta-learning principles can be useful to model animal adaptive behavior and underlying brain activity, (ii) in turn, when the machine

learning algorithms reach their limits, taking inspiration from known cognitive control mechanisms to improve these algorithms.

A simple heuristic proposed by Schweighofer and Doya (2003) consists in dynamically regulating RL parameters as a function of the agent's performance – *i.e.* as a function of the agent's current averaged obtained reward : at the beginning of the simulation, performance starts at a low level (the average reward is low), and thus the model starts with a high level of exploration ; while the agent progressively improves its performance, the exploration parameter is tuned so that there is less and less exploration ; as soon as a task change occurs, the average reward obtained by the agent drops, and thus the exploration parameter is reset to a high level. During my post-doctoral work in collaboration with Emmanuel Procyk and Peter F. Dominey, we have proposed a computational model for adaptive exploration regulation in the monkey prefrontal cortex (Khamassi et al. 2011a). We have shown that the model implemented on a humanoid robot can both (i) reproduce monkey performance in a problem-solving task with frequent task changes, (ii) enable the robot to display adaptive exploration regulation in an extended human-robot interaction game. The model was further used to draw a set of experimental predictions on prefrontal cortex activity that we later tested. The results are presented in a paper in press (Khamassi et al. 2014), included in Chapter 3.

However, this first stage of application of meta-learning principles to computational models of executive functions in primate relied on several simplifications, such as assumptions of reduced and stable environmental uncertainty. Further improvements could be done by taking inspiration from the way the brain uses volatility information to dynamically tune the learning rate parameter (Behrens et al. 2007) and performance at multiple-time scales to tune the discount factor (Tanaka et al. 2004).

Nevertheless, in addition to helping better understand brain functions, formalizing heuristics for the dynamical regulation of learning parameters and choice of the learning mode could be useful for Robotics, enabling robots to better cope with unexpected environmental changes and thus to display higher adaptivity and flexibility.

1.1.6 State of the art of decision-making and learning in Robotics

Major progress has been accomplished in several aspects of robotics : perception, navigation, localization, motion and action planning, manipulation, human-robot interaction (Siciliano and Khatib 2008). However most of the current results apply to restricted, pre-defined and well-known situations where robots' decisions only apply to quite simple problems. Moreover, robots learning abilities are still very limited, which requires the injection of prior knowledge by the human in the robot's decision-making system.

There have been applications of RL algorithms to robotics (*e.g.* Morimoto and Doya (2001), Smart and Kaelbling (2002), Alexander and Sporns (2002), Krichmar and Edelman (2002), Arleo et al. (2004)), some of which being neuro-inspired. But many of these studies – including ours (Khamassi et al. 2005; 2006) – produced limited progresses, due to applications to quite simple problems, with a small number of states and actions, to

slowness in learning and to systematic instability observed throughout the learning process. More recent applications of RL to robotics have permitted to deal with more complex and continuous action spaces, enabling to learn efficient sensori-motor primitives (Peters and Schaal 2008, Sigaud and Peters 2010, Kober and Peters 2011, Stulp and Sigaud 2013). But none of these approaches have attempted to equip robots with an ability to autonomously coordinate different learning systems through self-supervision and to decide which system should have the control over behavior at any given moment, as the mammalian brain does.

Besides, most of robotic decision-making algorithms are based on planning processes which take into account a great number of information, states, locations and actions (*e.g.* Chatila et al. (1992), Alami et al. (2006), Minguez et al. (2008), Kanoun et al. (2011)). Such approach to decision-making could be seen as similar to what we called the model-based system, except that there is most of the time no learning in the system : the internal model is given to the robot and only the planning, decision-making and execution parts have been addressed. Moreover, such an approach raises the issue of having to deal with high-dimensional state spaces, due to the combinatory explosion in large-scale applications. Another issue which is worthy of note is the long computation time imposed by the planning system, especially since there are systematic replanning of sequences of actions each time the robot is in the same situation and has to decide how to act. In contrast, mammals are able to use routines in familiar environments, controlled by their habit system which is in competition with their planning (model-based) system. The coordination of the planning system and low-level reactive routines is one the goals of cognitive architectures developed in Robotics (Alami et al. 1998, Volpe et al. 2001). Such architectures thus appear as a good direction of research for the coordination of decision-making systems in robots and to autonomously decide which system should take over the robot's behavior at each moment (Likhachev et al. 2002). Most of these architectures are built on the subsumption principle (Brooks 1986) in which different decisional layers are superposed in increasing order of complexity, each trying to control the robot's behavior, and superior layers being able to transiently take over (hence the term *subsumption*) inferior layers when it is appropriate. However, these architectures still lack efficient learning abilities and can thus not produce efficient behavioral adaptation in non-stationary environments.

Another field of robotics which is relevant for this HDR work and which in some cases include reinforcement learning algorithms is the study of robot navigation. In this paradigm, the objective of the agent is to reach a particular goal localized within the environment where the agent can receive a reward. The agent has to build a representation of space (*i.e.* a map enabling it to localize itself) and to learn how to reach the goal in the most efficient (quickest and safe) manner. In Robotics, map-based navigation in an a priori unknown environment is subject to several issues. First, in order to move along an appropriate trajectory, the system needs to autonomously build a relevant representation/map of the environment (which is the *mapping* step), to be able to know what is the robot's current location (*localization* step), and to be able to determine a path from point A to point B (*planning* step). While the *planning* step requires the other parts

to have been completed, the *localization* and *mapping* steps are mutually dependent : in order to localize oneself, it is necessary to recognize cues and features which characterize a particular place and which have previously been perceived and stored. Moreover, to build a reliable map and correctly situate features within it, the robot needs to be able to localize itself relative to these features (Angeli et al. 2008). These two steps thus have to be realized simultaneously, which gave the name to the central robotic problem of Simultaneous Localization and Mapping (*SLAM*) (Moutarlier and Chatila 1985).

The SLAM problem has been widely studied and numerous "engineering" solutions have been developed, in particular through three main paradigms : methods based on extended kalman filters (EKF, Smith et al. (1990)), graphical SLAM (Folkesson and Christensen 2004), and particule filters (Montemerlo et al. 2002). While numerous SLAM algorithms use the robot's laser to measure distances, there also exist SLAM algorithms based on camera vision. However, a major difficulty that SLAM algorithms face is the loop closure problem : recognizing places that the robot has already visited in order to obtain information which enables to correct estimation errors that have been accumulated with odometry. While some SLAM algorithms can work correctly without loop closure detection, the work of Angeli et al. (2008) shows that taking into account this aspect of navigation dramatically increases SLAM's results. Besides SLAM algorithms have difficulties to remain efficient when they are simulated for a long time. This is because the longer the robot navigates, the lesser true is becoming the hypothesis of a static world on which SLAM is anchored. Some promising solutions exist, for instance by using dynamical maps (Biber and Duckett 2005). But the issue is not solved yet. Finally, while SLAM algorithms focus on the localization and mapping aspects, they do not tell anything about how to make correct decisions using this information, nor how to adapt the robot's decisions through learning.

1.1.7 Neurorobotics approaches

Interestingly, several research groups have adopted a biomimetic approach to tackle some of these issues, with a two-fold objective : on the one hand, taking inspiration from the computational principles underlying mammals' behavioral flexibility to contribute to the improvement of current robots' autonomy and adaptivity (Frezza-Buet et al. 2001, Pfeifer et al. 2007, Meyer and Guillot 2008). On the other hand, using the robot as a platform to test the robustness of current biological hypotheses about cognitive functions, beyond perfectly controlled simulations, and try to learn more about the computational mechanisms at stake by analyzing which solutions enabled the model to work on a physical robot (Arbib et al. 2008).

In the particular case of robot navigation, several bio-inspired models of navigation have been tested in recent years, mostly inspired by rodent navigation. However, to our knowledge, none of these previous studies have addressed the issue of coordinating multiple decision and learning systems for navigation.

Arleo and Gerstner (2000) developed a computational model of place

cells - neurons located in the hippocampus whose activity encode an estimation of the animal's current position - and head-direction cells - neurons selective for the estimated orientation of the animal's head. With this model, they enabled a Khepera robot to navigate in a small arena, using a navigation strategy where learned associations between places and directions of movement – what is called a place-recognition triggerred response (*PRTR*) strategy and which can be learned with model-free RL. Fleischer, Krichmar and colleagues showed how prospective and retrospective coding at the level of place cells' activity can enable a robot to efficiently solve a spatial memory task (Krichmar et al. 2005, Fleischer et al. 2007); here also, navigation was performed by a PRTR strategy. Barrera and Weitzenfeld proposed a hybrid PRTR strategy using a graph, where the choice of the next action took into account the next three actions in a prospective manner (Barrera et al. 2011). Their robot could solve discretized implementations of various rodent laboratory mazes (T and radial mazes). Giovanangeli and Gaussier developed a model of another navigation strategy consisting in planning routes toward the goal in a topological graph ("cognitive map") of the environment – hence a "model-based" approach. Their model produced efficient navigation in both indoor and outdoor environments (Giovannangeli and Gaussier 2008). More recently, the RatSLAM algorithm has been implemented as a neural network inspired by the rat's hippocampus in order to perform efficient, continuous and long duration simultaneous localization and mapping (SLAM) on a robotic platform put in a large non-stationary environment (Milford and Wyeth 2010). Planning is also used here to perform navigation.

These different studies show efficient simulations of single navigation strategies, relying on a single learning system. The work presented in Chapter 4 shows how taking inspiration from mammals' ability to coordinate different navigation strategies, each equipped with specific learning mechanisms - namely *model-based* and *model-free* navigation strategies - can enable a robot to exploit the advantages of each strategy (work done by Ken Caluwaerts, a Master student that I co-supervised and whose work has already been mentioned above, Caluwaerts et al. (2012b;a)).

Other research groups have adopted similar bio-inspired approaches to study robotic cognitive functions or to more generally improve robots' adaptivity and autonomy. In particular, the Developmental Robotics approach attempts to mimick children's ability to learn sensori-motor affordances based on their intrinsic motivation to explore the environment (Lungarella et al. 2004, Oudeyer and Kaplan 2007). The Biomimetics approach concerns novel technologies developed through the transfer of function from biological systems (Lepora et al. 2013). In particular, this approach has made great advances in taking inspiration from animals' body properties and sensors which are not common in robotics, such as the rat's whiskers (Mitchinson et al. 2011, N'Guyen et al. 2011).

Based on this state of the art in neuro-inspired robotics, we further argue that incorporating bio-inspired meta-learning principles could enable robots to coordinate different learning systems through self-supervision in order to decide which system is the most efficient at a given time and in a given situation. This could help improve robots' flexibility and autonomy in decision-making.

1.2 Objectives and general approach

The main scientific issue addressed in this work is to understand how animals and robots can display behavioral adaptation capabilities in their partially unknown and changing environment. The objective is two-fold : on the one hand, contributing to better understanding behavioral and neural correlates of learning processes ; on the other hand, taking inspiration from biology to design autonomous robots able to learn from their own observations and errors.

As sketched in the introduction of the scientific context above, the work is built on previous evidence that the mammalian brain combines different memory systems which enable parallel learning processes for efficient behavioral adaptation – in particular processes called *model-based* and *model-free* Reinforcement Learning. Thus the goal of the work presented in this manuscript is to propose accurate computational models for the coordination of learning and decision-making systems observed in animals, and see whether these models can help better understand underlying brain activities as well as improving robots' adaptivity and autonomy.

1.2.1 Methodology and implementation

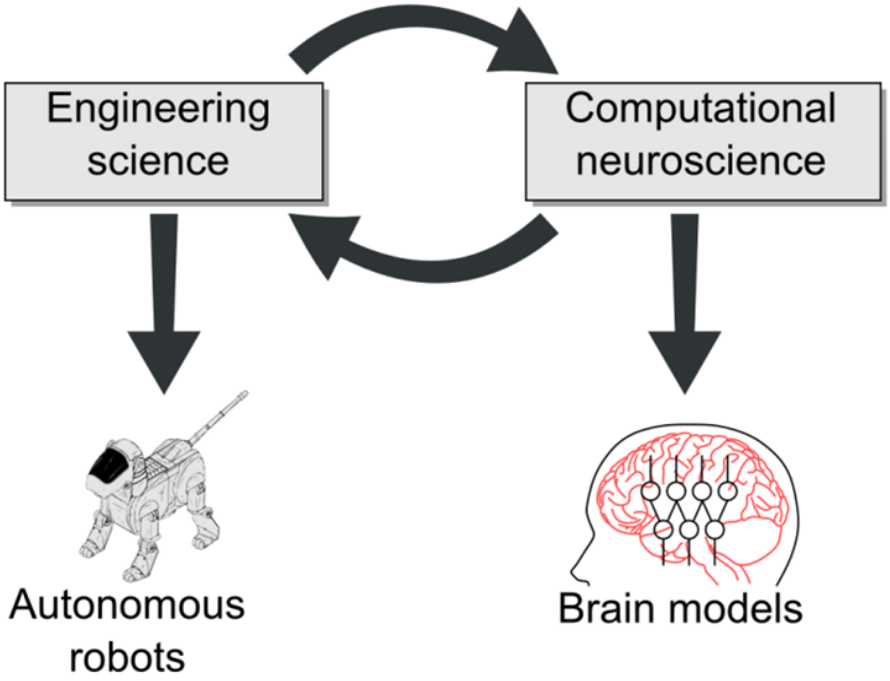


Figure 1.2 – *Illustration of the cross-disciplinary approach adopted in this research work (designed by Jean-Baptiste Mouret for the AMAC team at ISIR, UPMC-CNRS)*

The methodology implemented in this research work studying cognitive processes and their underlying neural structures is principally based on the conception of computational models. These models are then evaluated within the disembodied framework of Computational Neuroscience – *i.e.* comparison with electrophysiological, anatomical, behavioral data – and within the embodied framework of Cognitive Robotics – *i.e.* assessment of the efficiency of resulting controllers in the real world, com-

parison with engineering methods to evaluate the add-on of using neuro-inspired methods, new comparison with behavioral data and evaluation of the add-on compared to model simulations.

Such a research approach requires a strong interaction between Engineering Science and Computational Neuroscience (Fig. 1.2). The former brings machine learning algorithms, optimization tools, principles for robotic control systems. The latter brings the formalism and methodology of computational modelling, as well as methods for model comparison and model-based analyses on behavioral and neural data.

Although I have been trained to perform animal behavioral experiments and electrophysiological recordings – alongside computational modelling and robotic experiments – during my PhD work co-supervised by Sidney I. Wiener and Agnès Guillot (Khamassi 2007), the methodology adopted for this HDR research project does not include the realization of biological experiments myself anymore. It mostly relies on collaborations with experimentalists outside ISIR to design experiments enabling to address precise model predictions, to perform model-based analyses of biological data, and to extract principles from the results that can help improve computational models and robotic implementations.

1.2.2 Organization of the research work

Cross-disciplinarity interactions required for this research work are enabled within the *Architectures and Models for Adaptation and Cognition (AMAC)* team, coordinated by Stéphane Doncieux at the *Institute of Intelligent Systems and Robotics (ISIR, CNRS-UPMC)* and through external collaborations with experimentalists, theoreticians and roboticists.

The *AMAC* team gathers thirteen permanent researchers, with Computational Neuroscience, Computer Science and Robotics backgrounds, and is organized into five different research groups. I contribute to two of these groups : (i) the *Computational Neuroscience of Executive Functions* group in which Bruno Delord and Benoît Girard also participate ; (ii) the *Learning for Robotic Command and Decision-making* group in which Raja Chatila, Vincent Padois and Olivier Sigaud also participate. The local research environment of this work also includes a *LABEX* – a regrouping of research laboratories and institutes related to UPMC – called *SMART* and supported by French State funds managed by the ANR within the Investissements d'Avenir programme under reference ANR-11-IDEX-0004-02. Within this LABEX, I participate to one of the research programs aiming at modelling human learning abilities and I collaborate with the machine learning group of the Laboratory of Computer Science of UPMC (LIP6) with whom I co-supervise a PhD student (see Table 1.1).

Within this research environment, I currently co-supervise five PhD students (Table 1.1) – and have in addition and in total participated to the supervision of eight Master students, five Engineering students, and one external PhD student having performed a six-months research internship at ISIR. The firstly recruted PhD student, **Florian Lesaint**, has the goal of proposing a new multiple learning systems computational model accounting for behavioral phenomena involving the interaction between Pavlovian and Instrumental Conditioning, as well as dopamine activity

Name	Period	Main discipline	Co-supervision with
Florian Lesaint	2011–2014	Comp. Neuro.	Olivier Sigaud (ISIR)
Jean Bellot	2011–2014	Comp. Neuro.	Benoît Girard (ISIR)
Erwan Renaudo	2012–2015	Cog. Robot.	Raja Chatila (ISIR)
Nassim Aklil	2013–2016	Cog. Robot.	Ludovic Denoyer (LIP6)
Guillaume Viejo	2013–2016	Comp. Neuro.	Benoît Girard (ISIR)

TABLE 1.1 – *Co-supervised PhD students*

recorded by our collaborators during these experimental paradigms. The second PhD student, **Jean Bellot**, has the goal of proposing a new computational model of dopamine signalling in the basal ganglia and analyzing whether this model accounts for the information carried by dopamine neurons' activities recorded by our collaborators. The third PhD student, **Erwan Renaudo**, has the goal of implementing and testing a robotic architecture coordinating MB and MF reinforcement learning to enable robots to autonomously acquire behavioral habits, and see if the robot can constitute a good model of human habit learning in real-world continuous situations. The fourth PhD student, **Nassim Aklil**, has the goal of improving the coordination of learning systems within our current robotic multiple navigation strategy architecture with recent online budgeted learning techniques from the Machine Learning literature. The fifth PhD student, **Guillaume Viejo**, has the goal of proposing a new computational model for the coordination of learning systems to explain human behavior in tasks involving the interaction between reinforcement learning and working memory processes.

My research project is also made possible through external collaborations with experimentalists, theoreticians and roboticists, mostly in France, but also in other European Countries (United Kingdom, Italy, Switzerland, The Netherlands), in the United States of America, in Japan and in Tunisia. In particular, collaborators participating to the projects involving the PhD students I co-supervise or having contributed to the papers included in this manuscript comprise :

- The group of **Mark D. Humphries**, at Manchester University, UK, who designs computational models of action selection and performs model-based analyses of neurophysiological data (see Chapter 2).
- The groups of **Terry E. Robinson** and **Shelly B. Flagel**, at Michigan University, USA, who perform animal learning experiments, pharmacological manipulations and electrophysiological recordings during Pavlovian conditioning experiments (see Chapter 2).
- The group of **Kenji Doya**, at Okinawa Institute of Science and Technology, Japan, who performs animal learning experiments, computational models and robotics implementations of learning models (see perspectives in Chapter 2).
- The group of **Andrea Brovelli**, at CNRS in Marseille, who performs brain imaging in human experiments involving reinforcement learning, working memory and motor control processes (see perspectives in Chapter 2).
- The groups of **Emmanuel Procyk** and **Peter F. Dominey**, at INSERM in Lyon, who do electrophysiological recordings of monkey prefron-

tal cortex single-unit activity and local field potential during behavioral adaptation, and robotic implementations of neuromimetic models of cognitive functions (see Chapter 3).
- The groups of **Geoffrey Schoenbaum** and **Matthew R. Roesch**, at Maryland University, USA, who perform animal learning experiments, pharmacological manipulations and electrophysiological recordings during decision-making tasks (see Chapter 3).
- The group of **Etienne Coutureau** and **Alain Marchand**, at CNRS in Bordeaux, who do animal learning experiments and pharmacological manipulations during instrumental conditioning tasks (see perspectives in Chapter 3).
- The group of **Rachid Alami**, at CNRS in Toulouse, who works on shared action plans during human-robot interaction tasks (see perspectives in Chapter 4).
- The group of **Philippe Gaussier**, at Cergy-Pontoise University, who works on neuromimetic models of perception, navigation and social interaction (see perspectives in Chapter 4).
- The group of **Patrick Gallinari** and **Ludovic Denoyer**, at UPMC in Paris, who designs machine learning algorithms for large and structured dataset analyses with budget – *i.e.* computation time and cost – constraints (see perspectives in Chapter 4).

Outline of the presented work

Presentation of the research work is organized as follows :
- **Chapter 2** presents computational modelling work done to contribute in the formalization of principles underlying animals behavioral adaptation abilities. The work is presented under the form of two published journal papers (Khamassi and Humphries 2012, Lesaint et al. 2014). The first one has been performed with Mark D. Humphries and shows the relevance of using the model-based / model-free reinforcement learning computational framework to categorize navigation strategies in rodents and their underlying neural substrates. The second one presents the work of PhD student Florian Lesaint and shows that a computational model for the coordination of MB and MF RL enables to reproduce inter-individual behavioral and neurophysiological differences observed in rats called *sign-trackers* and *goal-trackers* in a Pavlovian conditioning paradigm.
- **Chapter 3** presents work employing the model-based analysis of neurophysiological data approach. The work is presented under the form of two journal papers, one in press (Khamassi et al. 2014), the other about to be submitted (Bellot et al. in preparation), aiming at testing model predictions about hypothesized neural activities underlying behavioral adaptation, and using the computational models to more precisely measure information related to particular computational mechanisms in neural activity. The first one has been performed with Emmanuel Procyk, Peter F. Dominey, René Quilodran and Pierre Enel and shows neural substrates of adaptive regulation of reinforcement learning parameters in the prefrontal cortical network during monkey behavioral adaptation. The second one presents the

work of PhD student Jean Bellot and shows model-based analyses of dopamine neurons' single-unit recordings during a decision-making task in rats.

– **Chapter 4** presents robotic implementations of neuro-inspired models of the coordination of MB and MF RL. The work is presented under the form of two papers, one published in a journal (Caluwaerts et al. 2012b), the other in the proceedings of an international conference (Renaudo et al. 2014), aiming at testing the ability of such neurocomputational models to improve robots' flexibility and adaptivity in real-world applications, and in return getting new insights into the properties of these computational models when tested in these more realistic conditions. The first one has been mainly performed by a previously supervised Master student, Ken Caluwaerts, and shows that the coordination of MB and MF learning systems for multiple-strategy-based navigation enables the robot to autonomously learn to exploit the advantages of each strategy in each subpart of the environment. The second one presents the work of PhD student Erwan Renaudo and shows that the coordination of MB and MF RL also enables to exploit the advantages of each system during a habit learning task in a humanoid robot. Both robotic studies shows that MB and MF systems do not behave exactly as expected by previous computational model simulations when they are interacting during embodied real-world applications.

Other published papers with supervised PhD students are not included in this manuscript, but will be discussed in relation to the presented work. These include (i) a paper presented at the Simulation of Adaptive Behavior Conference comparing the ability of different RL algorithms in reproducing dopamine activity (Bellot et al. 2012); (ii) a paper presented at the Living Machines Conference showing how extensions of the model-free learning system to take into account multiple landmarks within the environment can enable efficient coordination of MF and MB navigation strategies in a rat robot (Caluwaerts et al. 2012a); (iii) a paper submitted to a journal showing extensions of a MB / MF RL computational model to account for new Pavlovian conditioning data and draw a precise list of experimentally testable model predictions (Lesaint et al. submitted).

2 Computational modelling of the coordination of parallel learning processes in animals

Contents

This chapter presents computational modelling work done to contribute in the formalization of principles underlying animals behavioral adaptation abilities. The work is presented under the form of two published journal papers (Khamassi and Humphries 2012, Lesaint et al. 2014).

The first one has been performed with Mark D. Humphries and shows the relevance of using the model-based / model-free reinforcement learning computational framework to categorize navigation strategies in rodents and their underlying neural substrates. The proposed computational framework suggests that navigation strategies can be categorized as model-based or model-free, depending on the usage of information rather than on the type of information (*e.g.* cue versus place) as previous taxonomies propose. It moreover proposes that the Ventral Striatum (VS) participates to the model-building part of the involved computational processes.

The second one presents the work of PhD student Florian Lesaint and shows that a computational model for the coordination of MB and MF RL enables to reproduce inter-individual behavioral and neurophysiological differences observed in rats called *sign-trackers* and *goal-trackers* in a Pavlovian conditioning paradigm. The simulations suggest that the behavior of both types of animals is the result of a weighted sum of MB and MF

learning systems, with *sign-trackers*' behavior relying on a stronger weighting of the MF system while *goal-trackers*' behavior can be reproduced by a stronger weighting of the MB system. The model also explains why learning in *goal-trackers* has been experimentally shown to be dopamine-independent while this is not the case in *sign-trackers*.

2.1 Parallel navigation strategies

2.1.1 Khamassi and Humphries (2012) Frontiers in Behavioral Neuroscience

frontiers in
BEHAVIORAL NEUROSCIENCE

HYPOTHESIS AND THEORY ARTICLE
published: 27 November 2012
doi: 10.3389/fnbeh.2012.00079

Integrating cortico-limbic-basal ganglia architectures for learning model-based and model-free navigation strategies

***Mehdi Khamassi*[1,2]* *and Mark D. Humphries*[3,4]**

[1] *Institut des Systèmes Intelligents et de Robotique, Université Pierre et Marie Curie, Paris, France*
[2] *Centre National de la Recherche Scientifique, UMR7222, Paris, France*
[3] *Department d'Etudes Cognitives, Group for Neural Theory, Ecole Normale Superieure, Paris, France*
[4] *Faculty of Life Sciences, University of Manchester, Manchester, UK*

Edited by:
Matthijs Van Der Meer, University of Waterloo, Canada

Reviewed by:
A. David Redish, University of Minnesota, USA
Aaron Bornstein, New York University, USA
Hisham Atallah, Massachusetts Institute of Technology, USA

***Correspondence:**
Mehdi Khamassi, UPMC ISIR UMR 7222, Case courrier 173, 4 place Jussieu, 75005 Paris, France.
e-mail: mehdi.khamassi@isir.upmc.fr

Behavior in spatial navigation is often organized into map-based (place-driven) vs. map-free (cue-driven) strategies; behavior in operant conditioning research is often organized into goal-directed vs. habitual strategies. Here we attempt to unify the two. We review one powerful theory for distinct forms of learning during instrumental conditioning, namely model-based (maintaining a representation of the world) and model-free (reacting to immediate stimuli) learning algorithms. We extend these lines of argument to propose an alternative taxonomy for spatial navigation, showing how various previously identified strategies can be distinguished as "model-based" or "model-free" depending on the usage of information and not on the type of information (e.g., cue vs. place). We argue that identifying "model-free" learning with dorsolateral striatum and "model-based" learning with dorsomedial striatum could reconcile numerous conflicting results in the spatial navigation literature. From this perspective, we further propose that the ventral striatum plays key roles in the model-building process. We propose that the core of the ventral striatum is positioned to learn the probability of action selection for every transition between states of the world. We further review suggestions that the ventral striatal core and shell are positioned to act as "critics" contributing to the computation of a reward prediction error for model-free and model-based systems, respectively.

Keywords: reinforcement learning, habit, stimulus-response, action-outcome, nucleus accumbens

1. INTRODUCTION

A vast morass of neuroscience data addresses the problem of how voluntary behavior is underpinned by the anatomical and physiological substrates of the forebrain. Principles or frameworks to organize this data are essential. A consensus is growing around the potentially useful organizing principle that we can make a division of the forebrain striatum into three domains on both anatomical (Joel and Weiner, 1994, 2000; Voorn et al., 2004) and functional (Yin and Knowlton, 2006; Yin et al., 2008; Bornstein and Daw, 2011; Ito and Doya, 2011; van der Meer et al., 2012) grounds. From this "striatal eye-view" we can make sense of the wider cortical, hippocampal, amygdala, and basal ganglia networks in which they sit, and the role of these networks in different forms of voluntary behavior. Both the spatial navigation and instrumental conditioning literatures have adopted this perspective, recognizing the functional division of striatum into dorso-lateral (DLS), dorso-medial (DMS), and ventral striatum (VS) [1], belonging to different parallel cortico-basal ganglia loops (Alexander et al., 1990; Middleton and Strick, 2000), with each striatal domain having established functional roles within those broader behavioral distinctions. How do these functional distinctions map between the two literatures? And what might we learn by comparing the two?

While some links have been drawn between the approaches of the two literatures (Redish, 1999; Yin et al., 2004, 2008; Khamassi, 2007), their primary theories for the strategies underpinning behavior are, we suggest, orthogonal: the conditioning literature distinguishes *goal-directed* and *habitual* behavior in a task, whereas the navigation literature distinguishes *place* and *response* strategies for solving a task. However, there is mounting evidence that the place/response distinction is unable to account for the effects of lesions on navigation behavior. Our main hypothesis is that strategies for navigation, similar to strategies for instrumental conditioning (Daw et al., 2005), can be reconciled as either *model-free* or *model-based*—we define these terms below. At root, the key distinction is that it is the *use* of information in building a representation of the world, rather than the *type* of information about the world, that defines the different computational processes and their substrates in the striatum. We argue that explicitly identifying the DLS as a central substrate for model-free learning and expression, and the DMS as a central substrate for model-based learning and expression (Yin and Knowlton, 2006; Thorn et al., 2010; Bornstein and Daw, 2011; van der Meer et al., 2012) can help reconcile numerous conflicting results in the spatial navigation literature.

[1] We use VS throughout, rather than *nucleus accumbens*, to emphasize the contiguous nature of the striatum through its dorsolateral to ventro-medial extent (Voorn et al., 2004; Humphries and Prescott, 2010).

With this hypothesis in hand, we can see how work on spatial navigation gives us a second hypothesis, useful to understanding instrumental conditioning. We propose that the VS is a central substrate—in collaboration with the hippocampus—for a collection of functions that we informally term the "model-builder". On the one hand, the core of the VS acting as the locus of actions necessary to build a model; and on the other hand the shell of the VS acting to evaluate predicted and achieved outcomes in the model. These are clearly not the only roles of the multi-faceted VS (Humphries and Prescott, 2010); nonetheless, they may prove a further useful organizing principle.

With this sketch in mind, we address first the different forms of behavioral strategies that have separately been identified in the spatial navigation and instrumental conditioning literatures. We take a striatal-centric view here as an organizing principle, not as a claim that striatal domains are exclusive substrates for different forms of learning and navigation. Each striatal domain is one locus in a broader basal ganglia network that computes its output using information gathered by the striatum (Houk and Wise, 1995; Mink, 1996; Redgrave et al., 1999; Humphries et al., 2006; Leblois et al., 2006; Girard et al., 2008); and each network is in turn one locus in a broader basal ganglia-thalamo-cortical loop. Nonetheless, the striatum's consistent intrinsic microcircuit across the dorsolateral to ventro-medial axis (Bolam et al., 2006), its integration of cortical, thalamic, hippocampal, and amygdala input, and its position as the primary target of the midbrain dopaminergic system, makes it a natural vantage point from which to attempt to unify the disparate strands of navigation and conditioning.

2. STRATEGY DISTINCTIONS IN SPATIAL NAVIGATION

2.1. TAXONOMY OF SPATIAL NAVIGATION FORMS

Evidence for different navigation strategies in the rat comes from behavioral studies showing that they are able to rely on different information to localize themselves in the environment and to reach a certain location in space (Krech, 1932; Reynolds et al., 1957; O'Keefe and Nadel, 1978). Existing classifications of navigation strategies (O'Keefe and Nadel, 1978; Gallistel, 1990; Trullier et al., 1997; Redish, 1999; Franz and Mallot, 2000; Arleo and Rondi-Reig, 2007) point out a series of criteria, some of them overlapping, to differentiate navigation strategies: the type of information required (sensory, proprioceptive, internal), the reference frame (egocentric vs. allocentric), the type of memory at stake (procedural vs. declarative memory) and the time necessary to acquire each strategy (place-based strategies generally being more rapidly acquired than cue-guided strategies; Honzik, 1936; O'Keefe and Nadel, 1978; Packard and McGaugh, 1992, 1996; Redish, 1999). Moreover, it has been observed that in normal animals, a shift from a place strategy to a response strategy occurs in the course of training (Packard, 1999). This has led to the proposition of a strong distinction between two main categories of strategies:

- *Response* strategies, where a reactive behavior results from learning direct sensory-motor associations (like heading toward a visual cue or making an egocentric turn at the crossroads of a maze). This category includes target-approaching, guidance, cue-guided, and praxic [2] navigation, and can be further elaborated in the form of a sequence or chaining of Stimulus-Response (S-R) associations when new cues result from the previous displacement (O'Keefe and Nadel, 1978; Trullier et al., 1997; Arleo and Rondi-Reig, 2007).
- *Place* strategies, which rely on a spatial localization process, and can imply a topological or metric map of the environment (Tolman, 1948)—the term *map* being defined by Gallistel (1990) as "a record in the central nervous system of macroscopic geometric relations among surfaces in the environment used to plan movements through the environment".

2.2. SUBSTRATES IN THE STRIATUM

This strong strategy distinction has been mapped onto a strong distinction in underlying neural systems. It has been found that lesions of the hippocampal system impair place strategies while sparing response strategies (Morris, 1981; Packard et al., 1989; Devan and White, 1999). In contrast, lesions of the DLS produce the opposite effect: impairing or reducing the expression of response strategies while sparing place strategies (Potegal, 1972; Devan and White, 1999; Adams et al., 2001; Packard and Knowlton, 2002; Martel et al., 2007). Thus, it is common to speak of place and response strategies as being, respectively, "hippocampus-dependent" and "hippocampus-independent" (White and McDonald, 2002). Some theories propose that the "hippocampus-dependent" system expresses its output via the VS (Redish and Touretzky, 1997; Albertin et al., 2000; Arleo and Gerstner, 2000; Johnson and Redish, 2007; Penner and Mizumori, 2012). Other studies have also highlighted a role for the DMS in the "hippocampus-dependent" system (Whishaw et al., 1987; Devan and White, 1999; Yin and Knowlton, 2004), by finding that lesions of the DMS promote response strategies, implying the loss of place strategies. The behavioral strategies are often equated directly with learning systems: that is, separate systems that learn a particular cue-guided and/or place-guided set of strategies for a given environment. However, the simple mapping between VS-DMS vs. DLS onto place vs. response strategies is not consistent with mounting evidence from lesion studies.

2.3. KNOWN PROBLEMS WITH TAXONOMY AND SUBSTRATES

Response strategies are not solely dependent on the DLS. Chang and Gold (2004) reported that DLS-lesioned rats were only unable to express a response strategy on a T-maze in the absence of extra-maze cues; in cue-rich conditions the DLS-lesioned rats did not differ from controls in their ratio of using response or place strategies. Both Yin and Knowlton (2004) and De Leonibus et al. (2011) also found no significant decrease in the use of response strategies by DLS-lesioned rats running a T-maze. Moreover, Botreau and Gisquet-Verrier (2010) not only replicated this result but also ran a second separate cohort of DLS-lesioned rats to confirm it; further, they showed that the DLS-lesioned rats using a response strategy were really doing so: they continued to use that strategy to solve a new task on the T-maze.

[2] *praxic* normally refers to internally-generated sequences of movement independent of position information.

We conclude that the *response* learning system—including *cue-guided* and *praxic* strategies—cannot be simply associated with the DLS.

Place strategies are not solely dependent on the DMS. When learning to navigate to a hidden platform in the Morris water maze, rats with DMS lesions were able to learn the platform's location just as well as controls or DLS-lesioned rats, as indicated by their similar escape latencies (Whishaw et al., 1987; Devan and White, 1999); consistent impairment—shown by a lack of improvement over trials—only occurred if the fornix-fimbria [3] was cut (Devan and White, 1999). Botreau and Gisquet-Verrier (2010) reported that DMS-lesioned rats did not differ from controls or DLS-lesioned rats in their ratio of using response and place strategies in a probe test in the water-maze. We conclude that the *place* learning system cannot be simply associated with the DMS.

The precise role of VS in particular navigation strategies is even less clear (see Humphries and Prescott, 2010; Penner and Mizumori, 2012 for recent reviews). VS lesions impair place-based learning (Sutherland and Rodriguez, 1989; Ploeger et al., 1994; Setlow and McGaugh, 1998; Albertin et al., 2000). For instance, lesions of the medial shell of the VS impair the rat in learning and recalling the location of sites associated with larger rewards (Albertin et al., 2000). However, more recent studies reveal that VS function may not be restricted to place strategies. For instance, De Leonibus et al. (2005) report that VS lesions impair the acquisition of both allocentric and egocentric strategies in a task requiring the detection of a spatial change in the configuration of four objects placed in an arena.

The clean distinction between rapidly learnt place strategies and slowly learnt response strategies is also problematic. Several authors have reported rapidly learned response strategies (Pych et al., 2005; see Willingham (1998) and Hartley and Burgess (2005) for reviews including rodent data). Conversely, while place strategies have most of the time been found highly flexible and more rapidly acquired than response strategies (Packard and McGaugh, 1996), after extensive training place strategies can also become inflexible and persist in leading animals toward the previous goal location after a reversal, as if not relying on a cognitive map (Hannesson and Skelton, 1998; see also rat behavioral data in a Y-maze described in Khamassi, 2007).

These data suggest that the simple distinction between place vs. response strategies might be too broad to explain the different roles of VS-DMS vs. DLS in navigation. Several authors have highlighted that this classification of navigation strategies lends too much importance to the *type* of information involved (i.e., place vs. cue) and thus to the spatial localization process (Trullier et al., 1997; Sutherland and Hamilton, 2004). We suggest that considering the type of learning involved—and measurable in terms of behavioral flexibility—might better account for the specific involvement of VS, DMS, or DLS in navigation. To see this, let us first consider the taxonomy of learning in instrumental conditioning.

[3]This fiber pathway brings hippocampal information to the VS, but is also the source of brainstem inputs to the hippocampus, so may disrupt either transmission of place information by hippocampus or the encoding of place in hippocampus.

3. STRATEGY DISTINCTIONS IN INSTRUMENTAL CONDITIONING

3.1. GOAL-DIRECTED BEHAVIORS vs. HABITS

A long line of conditioning research has elaborated two operationally defined forms of instrumental behavior in the rat: *goal-directed* in which the animal is able to modify its behavior in response to changes in outcome and *habitual* in which the animal does not respond to changes in outcome (it perseveres with its previous action— hence "habit") (Dickinson, 1985; Yin et al., 2008). This definition is "operational" because it can only be safely defined in retrospect— i.e., after extinction. Experimenters typically use a test in extinction to discriminate between these two behavioral modes after a reward devaluation or change in contingency between behavior and reward. If during this extinction test the animal quickly stops producing the now irrelevant conditioned response (e.g., pressing a lever) it is said to be *goal-directed*; if the animal persists it is said to be *habitual* (Balleine and Dickinson, 1998). The inference is then drawn that goal-directed animals have access to action-outcome contingencies to guide behavioral choice, and that changes in outcome consequently change action choice, whereas habitual animals make behavioral choices based on S-R pairings (Dickinson, 1985).

3.2. SUBSTRATE EVIDENCE FOR DMS' GOAL-DIRECTED AND DLS' HABITUAL ROLES IN LEARNING

During the course of a conditioning task animals' behavior progressively shifts from expressing awareness of action-outcome contingencies to expressing habits. In particular, after extensive training or *overtraining* animals' behavior is most often habitual (Yin et al., 2004). It turns out that this natural progressive shift can be perturbed by lesions of different parts of the striatum, pointing to a possible double-dissociation between DLS and DMS: the former being required for acquisition and maintenance of habits, and the latter being required for learning and expression of goal-directed behaviors (Balleine, 2005; Yin and Knowlton, 2006; Yin et al., 2008).

There is a strong consensus that the dorsolateral striatum is necessary for habitual behavior: lesions of either the DLS (Yin et al., 2004), or disruption of dopamine signaling within it (Faure et al., 2005), prevent habit formation in extinction. Animals with such lesions thus appear to maintain goal-directed behavior throughout a task. Correspondingly, there is a re-organization of the DLS' single neuron activity during habit formation (Barnes et al., 2005; Tang et al., 2007; Kimchi et al., 2009). Consequently, the dorsolateral striatum has been proposed as central to the learning of habits (Yin and Knowlton, 2006; Yin et al., 2008).

There is a strong consensus that the dorsomedial striatum is necessary for goal-directed behavior: lesions of the DMS (Yin et al., 2005b), or blockade of NMDA receptors within it (Yin et al., 2005a), putatively preventing synaptic plasticity, prevent sensitivity to devaluation or contingency changes in extinction. Animals with such lesions thus appear to obtain habitual behavior from the outset. Correspondingly, there is a re-organization of the DMS' single neuron activity after changes in action-outcome

contingencies (Kimchi and Laubach, 2009; Kimchi et al., 2009). Consequently, the dorsomedial striatum has been proposed as central to goal-directed learning (Yin and Knowlton, 2006; Yin et al., 2008).

A caveat is that the anterior part of DMS (aDMS) may escape from this functional scheme. To our knowledge, only the posterior DMS (pDMS) has been clearly shown as involved in the acquisition of goal-directed behaviors (Yin et al., 2005b) and in place-based navigation (Yin and Knowlton, 2004). Lesions of aDMS do not affect either of these processes. They even increase the number of rats classified as place-responders both during initial and late phases of learning (Yin and Knowlton, 2004), and seem to increase the sensitivity to contingency degradation (compared to sham-lesioned rats) (Yin et al., 2005b). Ragozzino and Choi (2004) showed that inactivating aDMS does not affect learning of a T-maze task or acquisition of a place strategy; but inactivation during reversal learning did affect performance, thus suggesting that aDMS is involved in switching between strategies, not in learning *per se*. Contrary to these data, Moussa et al. (2011) showed that a rat's impairment in learning an alternating-arm T-maze task correlated with volume of DMS damage, not with the location of the lesion. Nonetheless, it remains possible that the aDMS is not part of the goal-directed or habitual systems.

3.3. THE VENTRAL STRIATUM IN CONDITIONING

While dorsal parts of the striatum are important for the expression of learned S-R contingencies, their acquisition may require intact VS (Atallah et al., 2007). The VS is indeed located at a crossroads between limbic and motor structures which places it in a privileged position to integrate reward, motivation, and action (Mogenson et al., 1980; Groenewegen et al., 1996). In the instrumental conditioning literature, the VS is also considered particularly important for Pavlovian influences over voluntary behavior (Balleine and Killcross, 1994; Dayan and Balleine, 2002; Yin et al., 2008; van der Meer and Redish, 2011). It has been attributed roles as both a locus of Pavlovian conditioning—learning to associate outcomes to different stimuli or states—and the locus of Pavlovian-instrumental transfer—the use of those learnt stimulus-outcome associations to motivate the learning and expression of instrumental actions in the presence of those stimuli (Yin et al., 2008). Further, while the functional subdivision of VS into core and shell might be oversimplified (Heimer et al., 1997; Ikemoto, 2002; Voorn et al., 2004; Humphries and Prescott, 2010), it may account for distinct influences of reward values on habitual performance and goal-directed behavior, respectively. For instance, Corbit and Balleine (2011) found that shell lesions impair outcome-specific [putatively goal-directed as noted by Bornstein and Daw (2011)] Pavlovian-instrumental transfer while core lesions impair general (putatively habitual) Pavlovian-instrumental transfer.

These data suggest that the differences in the learning process controlling the progressive influence of rewards on actions may determine the functional roles of striatal domains in various behavioral strategies: DLS being involved in learning and expression of habitual behaviors; DMS being involved in learning and expression of goal-directed behaviors; VS controlling the influence of reward values on these two processes during learning. Computational work has brought great advances in formalizing the differences between these learning processes.

3.4. MODEL-BASED vs. MODEL-FREE LEARNING PROCESSES

Machine-learning research into formal algorithms for reinforcement learning has developed a basic distinction between two forms of such algorithms. Common to both is the idea that we can represent the world as a set of states $\mathbb{S}$, that the agent could take one of a set of actions $\mathbb{A}$ in each state (including no action at all), and that the outcome of taking action a in state s is the next state s' and a possible reward r (Sutton and Barto, 1998). Distinguishing the two is whether or not the dependencies in the world representation are explicitly modeled (**Figure 1**).

In the *model-free* forms of algorithm, each state has associated with it a distribution of the values of each possible action, learnt iteratively using a prediction error to minimize the difference between the values of actions in consecutive states. This set includes most well-known forms of reinforcement learning algorithms—including Temporal Difference (TD) learning, Actor-Critic, and Q-Learning. Each state thus has an associated distribution of cached action-values $Q(s, a)$ over all available actions. The action to execute is then simply chosen based on this cached value distribution. Such behavior is called reactive in that it is state-driven—e.g., stimulus-driven—and does not rely on the inference of possible outcomes of the action.

In the *model-based* forms of algorithm, direct use is made of the state information about the world. With each state s is still associated a reward r, each action is still assigned a value $Q(s, a)$, and action selection is based on those values. However, model-based algorithms explicitly store the state transitions after each action: they can then simulate off-line the consequence of action choices on transitions between states before choosing the next action appropriately (Sutton and Barto, 1998; Johnson and Redish, 2005). Thus in this case the agent will infer possible future outcomes of its decisions before acting. In simple decision-making tasks in which each action leads to a different state, such a process is naturally captured by a branching decision tree (**Figure 1**); in more natural situations states may be re-visited during ongoing behavior, and thus the transitions between states may have periodic structure. Sophisticated model-based algorithms explicitly compute a separate transition matrix $T(s', a, s)$ for the probability of ending up in each next state s', given the current state s and each possible action choice a in $\mathbb{A}$ (Daw et al., 2005, 2011; Glascher et al., 2010).

Daw et al. (2005) proposed the formal mapping that goal-directed behavior results from model-based learning and that habitual behavior results from model-free learning[4]. They further proposed that both learning systems operate in parallel, with

[4]They used a model-based algorithm that explicitly computed the transition matrix. It seems feasible that simpler model-based algorithms, without explicit computation of the transition matrix, could also equally account for the sensitivity to devaluation and contingency changes in goal-directed learning, as their repeated internal simulation after such outcome manipulations would result in more rapid changes in overt behavior. To our knowledge, no one has examined the possibility. Intriguingly, Johnson and Redish (2005) showed that such an internal-simulation model, emulating hippocampal

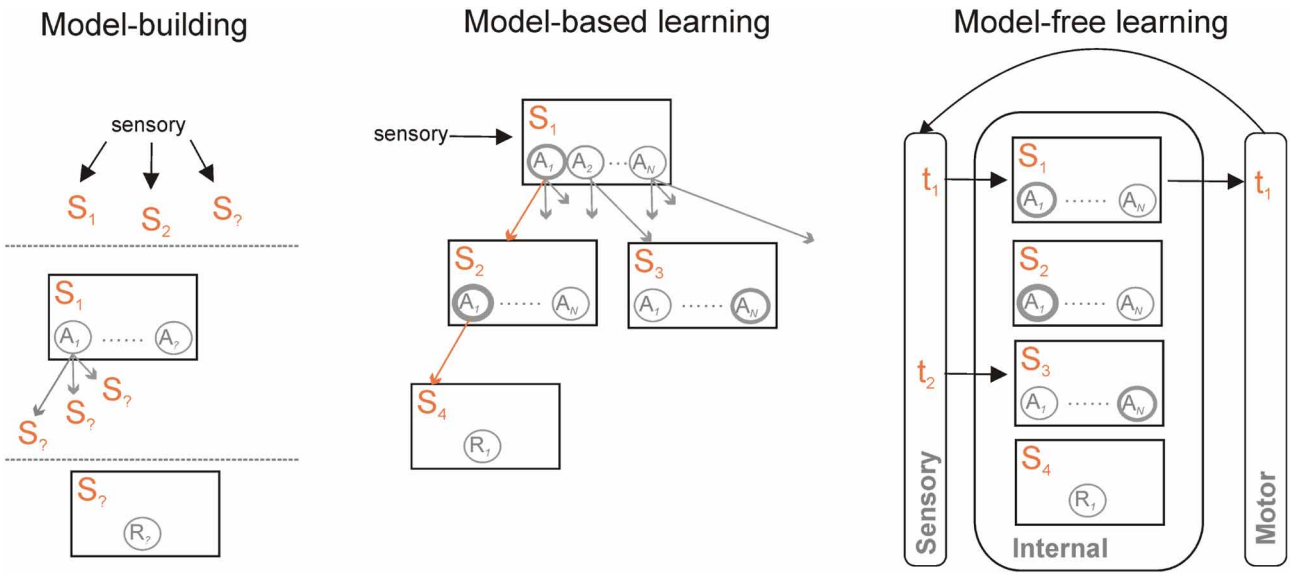


FIGURE 1 | Model-based and model-free learning and controllers. Model-based and model-free controllers represent the world as a set of states $S_1 \ldots S_m$ and actions $A_1 \ldots A_N$ within those states. They learn the values of each action in a given state, here indicated by the thickness of each circle, based on available rewards R. What distinguishes them is their representation of the links between those states. A model-based controller **(centre)** also represents the transitions between states and the action(s) that cause the transition (indicated by the multiple arrows). For a known current state, specified by current sensory information, the model can be traversed to find the likely outcome of simulated actions in each state—one such trajectory is given by the orange arrows. Each trajectory can then be used to update the predicted value of each action. Finally, after a number of trajectories through the model, an overt action is selected based on their updated values in the current state. A model-free controller **(right)** vastly reduces the representational and computational demands by essentially externalizing the world-model. Sensory information specifies the state at time t_1; an action is chosen based on its current value. Updated sensory information resulting from that action then specifies the state at time t_2. Learning is then based on the prediction error between expected and resulting values of the action taken at t_1. A model-free controller can also be trained by a model-based controller, and thus represent an abstraction of that model. Irrespective of whether model-free or model-based, a common set of information needs to be learnt to construct and use the controller **(left)** to specify the set of current relevant states in the world; to learn actions available within them and the transitions those actions cause; and to learn the reward function—which state(s) contain reward(s).

the system chosen for current behavioral control based on having the least uncertainty in its prediction of the outcome. Using stylized examples of simple conditioning tasks, they showed how this mapping can explain the sensitivity to devaluation and contingency degradation in extinction early in training when the model-based controller is dominant, and how that sensitivity is lost when the model-free controller becomes dominant with over-training. The underlying explanation is that the model-based controller directly represents action-outcome contingencies, and is thus able to quickly propagate changes in reward through the world-model; by contrast, the model-free controller, while able to reduce the uncertainty in its predictions with over-training, requires further extensive training for the change in reward to propagate through the independent state-action representations. This formal mapping onto computational substrates has proven a very useful and fruitful guide to the understanding of these operationally-defined forms of behavior and their inferred learning systems (Ito and Doya, 2011; Bornstein and Daw, 2011; van der Meer et al., 2012).

This computational mapping is also assumed to follow the same substrate mapping (Daw et al., 2005; Bornstein and Daw, 2011; Ito and Doya, 2011). Thus, as DLS is central to the habit-learning system, so, by extension, it is considered central to the model-free learning system in instrumental conditioning (Daw et al., 2005). Similarly, as DMS is central to the goal-directed system, it is thus natural to propose that DMS is central to the model-based learning system in instrumental conditioning (Bornstein and Daw, 2011).

4. UNIFICATION: NAVIGATION STRATEGIES ARE MODEL-FREE OR MODEL-BASED

Superficially, the model-free/model-based dichotomy strongly resembles the dichotomous taxonomy defined in the spatial navigation literature between flexible map-based *place* strategies and automatic map-free *response* strategies. However, the two approaches are orthogonal: one is defined by information use in a world representation (model-free/based), the other by information type (place/cue).

Our hypothesis is that we may similarly distinguish model-free and model-based navigation strategies by their use of information (**Figure 2**), no matter if the state is represented by a spatial location or a visual stimulus. Within these two top-level strategies, we may further differentiate strategies defined by their reference frame and modality of processed stimuli:

replay of previous trajectories through a maze, could indeed reduce the onset of habit-like stereotypy in the paths taken through the maze.

A

		Action selection process	
		Inflexible, slow to acquire (S-R associations)	Flexible, rapidly learned (cognitive graph)
Strategy dimension	Place		**Place strategies**
	Cue	**Response strategies**	

B

		Action selection process	
		Inflexible, slow to acquire (model-free)	Flexible, rapidly learned (model-based)
Strategy dimension	Place	**Place strategies** (PRTR)	**Place strategies** (map-based)
	Cue	**Response strategies** (habitual)	**Response strategies** (goal-directed)

FIGURE 2 | New taxonomy of navigation strategies based on model-based/model-free reinforcement learning. (A) Previous taxonomies highlight the distinction between flexible rapidly acquired map-based strategies and inflexible slowly acquired S-R strategies. **(B)** New taxonomy highlighting model-free and model-based place strategies as well as model-free and model-based response strategies. PRTR, place-recognition triggered response strategies as classified by Trullier et al. (1997).

- egocentric reference frame, relying on idiothetic (praxic), or allothetic (cue-guided) stimuli;
- allocentric reference frame, relying on idiothetic and/or allothetic stimuli (places).

Our hypothesis thus naturally extends to proposals for the striatal substrates of model-free and model-based strategies in navigation: that the DLS is central to the model-free navigation system and DMS is central to the model-based navigation system.

This combined conceptual (model-free vs. model-based) and substrate (DLS vs DMS) hypothesis raises four implications that each explain some troubling or inconsistent data for the place vs. response dichotomy in navigation. First, that we can conceive of a model-free strategy based on place information alone supported by the DLS. Second, that, correspondingly, we can conceive of a model-based "response" strategy based on cues alone supported by the DMS. Third, that, following the model-based/model-free mapping in conditioning (Daw et al., 2005), model-based and model-free control of navigation could be distinguished behaviorally by whether or not the animal reacts to changes in the value or contingencies of rewards, and by lesions to the DLS and DMS. Fourth, that both place and cue information should be available to both the model-based and model-free navigation systems, and thus should be detectable within both the DMS and DLS. We consider each of these in turn, then discuss the key role of the hippocampal formation as the likely source of state information.

4.1. DLS AND (MODEL-FREE) PLACE STRATEGIES

Model-free navigation strategies based on place information alone have been called "Place-Recognition Triggered Response (PRTR)" strategies by Trullier et al. (1997) who emphasized that such a strategy produces inflexible behavior because it needs to relearn sequences of place-response associations in case of a change in goal location. This type of learning was prominent in early models of hippocampus-dependent navigation (Burgess et al., 1994; Brown and Sharp, 1995; Arleo and Gerstner, 2000; Foster et al., 2000).

Following the same DLS vs. DMS double-dissociation logic as was used for goal-directed and habitual learning then, if DMS is the substrate for place strategies, lesions of the DMS should impair place strategies and lesions of the DLS should not affect them. However, there is evidence against this dissociation and indirect evidence in favor of a place strategy supported by DLS. Lesions of the DMS slow but do not prevent the learning of a hidden platform in a water maze, which putatively requires a place-based strategy (Devan and White, 1999). More compelling, Botreau and Gisquet-Verrier (2010) tested control, DLS-lesioned, and DMS-lesioned rats learning a hidden platform water maze task; after learning, a probe trial was used where the rats were started in a different location for the first time: they found that rats were divided into the same ratio of "place" and "response" groups on the probe trial irrespective of whether they were control, DLS-lesioned, or DMS-lesioned rats. Recently, Jacobson et al. (2012) tested rats on an alternating strategy plus-maze, which required the use of either a response-based or place-based strategy on each trial as signaled by an extra-maze cue: they found that post-training DLS lesions impaired use of both the response and place strategies. Thus, there is evidence that intact DLS is important for using place strategies.

4.2. DMS AND (MODEL-BASED) RESPONSE STRATEGIES

The proposal of a model-based response strategy is just the claim that we can conceive of states in a spatial navigation task as being defined by the position of intra- or extra-maze cues relative to the animal. In such a model, different states would not necessarily correspond to different spatial position. Rather, we can conceive of an example task where distinct states s_1 and s_2 correspond to the same spatial location and differ on whether a light is turned on or off. Then a model-based system can learn the transitions between these states and search the model to proceed with action selection—e.g., reward may be delivered only when the light is on. Thus, whereas others have explicitly identified a response strategy—e.g., a strategy guided by the light—with habitual behavior (e.g., Yin and Knowlton, 2004), we are proposing that the two are orthogonal.

Again we may follow the same double-dissociation logic: if DLS is the sole substrate for response strategies, then lesions of the DLS should impair response strategies and lesions of the DMS should not affect them. There is evidence against this dissociation, and in favor of DMS involvement in response-strategies. As noted in section 2.3, lesions of the DLS do not impair the use of response strategies on probe trials, suggesting that intact DMS is sufficient to support the use of response strategies (Chang and Gold, 2004; Yin and Knowlton, 2004; Botreau and Gisquet-Verrier, 2010; De Leonibus et al., 2011). Chang and Gold (2004) further reported that the DLS lesions only effectively impaired the use of response strategies when there were no extra-maze cues. This suggests that model-based (and putatively DMS-based) use

of cues was sufficient to maintain a response strategy in the cue-rich conditions; but that a model-free (and putatively DLS-based) praxic response strategy was necessary in the cue-deficient conditions (that is, in the absence of sufficient cues, learning a sequence of turns was required).

Moussa et al. (2011) tested the effects of DLS and DMS lesions on the ability of rats to learn a return-arm T-maze in which the rats were required to alternate their choice of visited arm (left or right) to obtain reward, but were free to run at their own pace. The task is a seemingly simple response strategy but requires a minimal model to achieve rewards above chance level. At the choice point of the T-maze, a model-free learning system would assign equal value to turning left or turning right as both would be rewarded on (approximately) half the visits. To achieve better, a minimal model would be needed to at least link the previous choice of arm to the current choice, chaining at least two (state, action) pairs in a loop—which corresponds to a model-based process. Moussa et al. (2011) found that DMS lesions, and not DLS lesions, impaired learning of this task irrespective of the amount of training. Their data thus suggest a model-based response strategy role for DMS.

4.3. VALUE-SENSITIVITY IN NAVIGATION AND ITS ALTERATION BY DMS BUT NOT DLS LESIONS

If the prediction of Daw et al. (2005) is correct, then model-based and model-free control of action can be distinguished behaviorally by whether or not the animal reacts to changes in the value or contingencies of rewards. Thus, under our hypothesis, such sensitivity to value or contingency changes in spatial navigation should be reflected in *both* place and response strategies if using a model-based controller and in *neither* place nor response strategy if using a model-free controller. Similar to the goal-directed to habitual transfer observed in instrumental conditioning (Yin and Knowlton, 2006), we might expect that this outcome sensitivity would disappear with over-training on a sufficiently deterministic task, reflecting the transfer from a model-based to a model-free controller for navigation. Also similarly, our hypothesis is that this transfer is from the DMS to the DLS-based systems; so lesions to those systems should differentially affect how changes in value subsequently change behavior.

Whereas above we reviewed evidence in favor of their breaking the place vs response dichotomy, here we consider evidence more directly in favor of the association of DMS with a model-based system and DLS with a model-free system. De Leonibus et al. (2011) recently provided intriguing evidence from devaluation in favor of both (1) the existence of model-based and model-free response strategies and (2) their dissociable modulation by DMS and DLS lesions. Further, Moussa et al. (2011) provided evidence from extinction during navigation for both. We consider these studies in turn.

Figures 3A,**B** outlines De Leonibus et al. (2011) dual-solution plus-maze task and experimental design. Key to the design was separately training "early" and "late" groups of rats for, respectively, 26 and 61 days before the first probe trial, which established the strategy they were using to locate the reward (**Figure 3B**). Both "early" and "late" groups preferentially used the response strategy on the first probe trial (**Figures 3C**,**F**), replicating earlier results (Devan and White, 1999; Yin and Knowlton, 2004). However, the response strategy sub-group for both "early" and "late" were then split, with approximately half receiving a devaluation regime for the food reward in the maze. On the subsequent second probe trial, only the "early" group showed awareness of the devaluation, through a significant drop in their use of a response strategy (**Figure 3D**). There was no change in the use of response strategy by the devalued "late" group (**Figure 3G**). Thus, while both "early" and "late" groups of rats preferentially used a response strategy, only the early group modified use of that strategy after change in the value of reward, evidence of a distinction between a model-based and model-free form of response strategy.

De Leonibus et al. (2011) then separately tested the effects of pre-training sham and DMS lesions on a new "early" group, and of pre-training sham and DLS lesions on a new "late" group. They found that the DMS lesion prevented the devaluation from changing the proportion of "early" group rats using a response strategy (**Figure 3E**). This is consistent with the loss of DMS preventing value updates from propagating through the model-based system. Conversely, they found that the DLS lesion now permitted the devaluation to change the proportion of "late" group rats using a response strategy (**Figure 3H**). This is consistent with the loss of DLS preventing transfer to the model-free system, and subsequently value updates continued to propagate through the model-based system. Together, these results support the double dissociation of DMS as part of a model-based and DLS as part of a model-free system for navigation.

Moussa et al. (2011) found results consistent with this picture from rats tested in extinction on a navigation task. As noted above, they tested rats on an alternating arm T-maze task, thus requiring rats to maintain a memory of the previously visited arm. As the rats ran at their own pace, Moussa et al. (2011) were unusually also able to test the effects of extinction on navigation tasks by leaving the arms unbaited in the final 10-min session. They found that control rats did decrease their laps of the maze over the 10-min period, so that extinction effects were detectable. Moreover, though DLS lesions had no effect on learning the task, they did lead to significantly faster extinction of maze running. These data are thus consistent with lesions of DLS removing the putative model-free navigation substrate, thus leaving intact the putative model-based substrate in DMS that was subsequently faster to respond to the outcome devaluation.

4.4. PLACE AND CUE INFORMATION IS AVAILABLE TO BOTH MODEL-BASED AND MODEL-FREE SYSTEMS

If the DLS and DMS are indeed, respectively, substrates for model-free and model-based navigation systems, and not the response and place systems, then cue- and place-based correlates of movement should appear in the activity of both.

DLS activity is consistent with the development of cue-based correlates of movement. Jog et al. (1999) showed that developing DLS activity over the course of a T-maze task stabilized to just the start and end positions in the maze once the rats had reached operationally "habitual" behavior. van der Meer et al. (2010) showed that decoding of position information from dorsal striatal activity consistently improved over experience, and that its

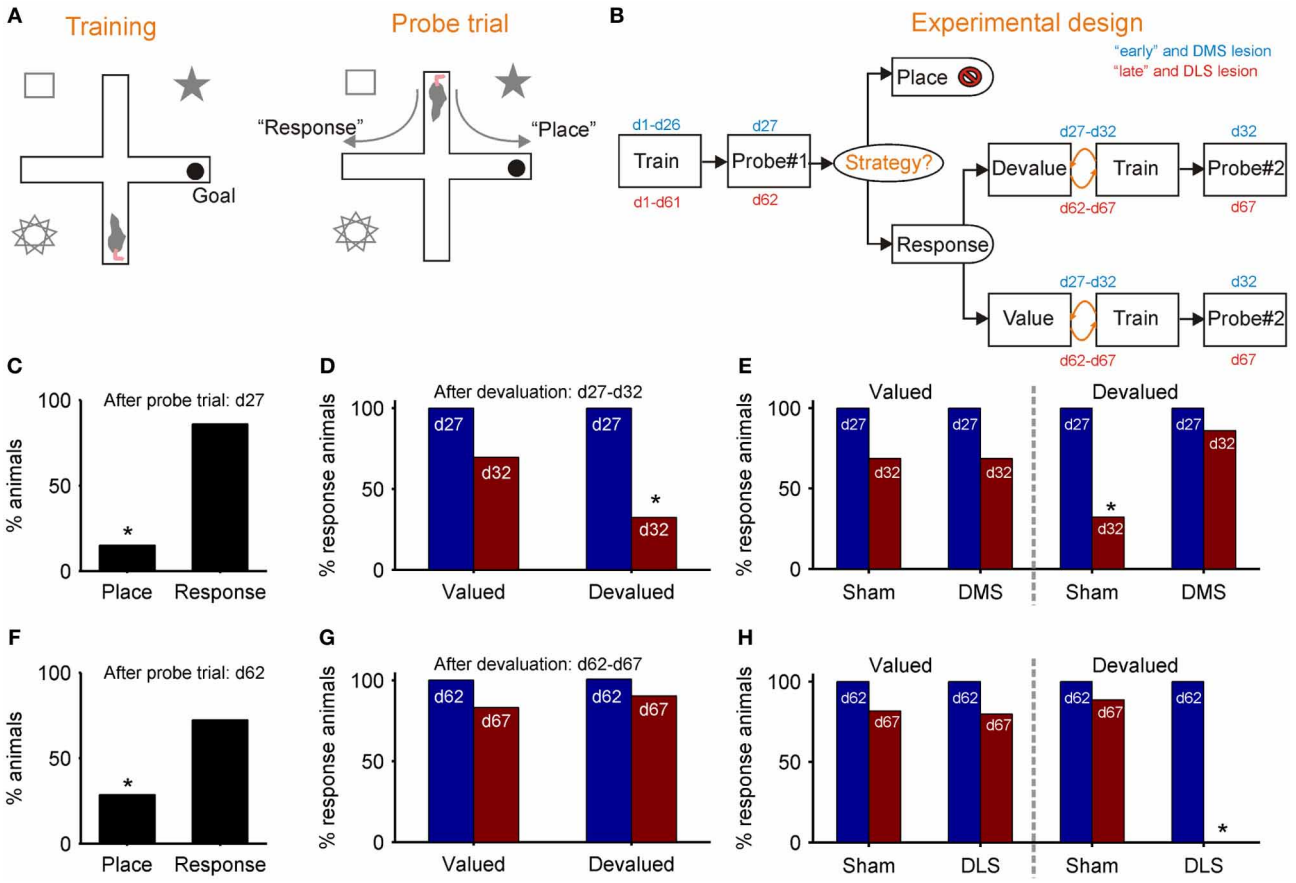


FIGURE 3 | Evidence for model-based and model-free navigation in data reported by De Leonibus et al. (2011). (A) Dual-solution plus-maze task used by De Leonibus et al. (2011). On training trials, rats always start from the same arm (south) and have to learn the location of the reward in a consistently baited arm (e.g., east). After training, a probe trial starting in the opposite arm is used to ascertain the rat's strategy for locating the reward (a food pellet): a "response" strategy based on direction of turn, or a "place" strategy based on location of reward with respect to extra-maze cues. **(B)** The experimental design of De Leonibus et al. (2011). Rats were in two broad categories, designated "early" and "late" with respect to the first probe trial (day 27 or day 62). All "response" rats from that trial were taken forward to the second stage, and split approximately evenly into devaluation and control (value) groups. Both groups had free access to food pellet reward for 15 min immediately after training for each of five days; the devaluation group received an injection of LiCl immediately afterwards, the control group received a saline injection. The devaluation group developed a taste aversion to the pellets, but no reduction in completed trials (De Leonibus et al., 2011). **(C–E)**: data from "early" group; **(F–H)** data from the "late" group. **(C)** Proportion of "early" group rats using each strategy on first probe trial. **(D)** From the second probe trial, the proportion of rats continuing to use a "response" strategy after devaluation compared to controls. **(E)** From the second probe trial, the proportion of rats continuing to use a response strategy after devaluation and pre-training DMS lesion, compared to controls for both. **(F)** Proportion of "late" group rats using each strategy on first probe trial. **(G)** From the second probe trial, the proportion of "late" group rats continuing to use a "response" strategy after devaluation, compared to controls. **(H)** From the second probe trial, the proportion of rats continuing to use a response strategy after devaluation and pre-training DLS lesion, compared to controls for both. An $*$ indicates a significant difference of at least $p < 0.05$—see De Leonibus et al. (2011) for details.

activity peaked only at choice points in the maze, consistent with a slow learning model-free system that learnt to associate differentiable intra-maze states with actions (Graybiel, 1998; Yin and Knowlton, 2006). DLS activity is also selectively correlated with position: Schmitzer-Torbert and Redish (2008) found that dorsolateral striatal electrophysiological activity correlated with place when the task required knowledge of spatial relationships, but no correlation when the task was non-spatial.

DMS is clearly in receipt of place information in that activity is correlated with actions or rewards in particular locations, but not correlated with the location alone (Wiener, 1993; Berke et al., 2009). Furthermore, lesions of posterior DMS prevent execution of place-based strategies (Yin and Knowlton, 2004) as does loss of dopamine from that region (Lex et al., 2011). Its input from the prefrontal cortex (PFC), particularly medial PFC which receives considerable direct input from the CA1 place cells, is one of the most likely sources of place information; there is clear evidence that medial PFC supports place representation [e.g., Hok et al. (2005)]. Nonetheless, there is also evidence for DMS' receipt of cue-information. Devan and White (1999) reported that asymmetric lesions (unilateral hippocampus and contralateral DMS) produced mild retardation of acquisition of both cue-based and place-based learning. Correspondingly, recording studies report that the largest changes in DMS neural activity occur in the

middle stages of learning during cue-guided (both with auditory and tactile cues) navigation (Thorn et al., 2010).

4.5. HIPPOCAMPAL INPUT TO MODEL-BASED AND MODEL-FREE SYSTEMS

For spatial navigation the primary candidate for generating the states and the relationship between them is the hippocampal formation. Although hippocampus has been largely associated with spatial encoding (O'Keefe and Nadel, 1978), it could be more broadly involved in learning (and planning in) a model or graph of possible transitions between states, no matter if these states are spatial or not (van der Meer et al., 2012). Consistent with this, hippocampal place cells are also sensitive to non-spatial information (e.g., the presence of a certain object or the color of the walls), this non-spatial information modulating or re-mapping the place representation (Wiener et al., 1989; Redish, 1999). Similarly, hippocampal place cells re-map on maze tasks following a change of context, such as the change of rewarded arm in a plus-maze (Smith and Mizumori, 2006). Thus, within our proposal, the role of the hippocampus would be both to supply spatial information to a model-free system and to contribute to a model-based system by building the model—in interaction with the VS as argued later—and planning actions within this model. This view is similar to ideas that the hippocampus provides contextual information to some aspects of learning such as contextual fear conditioning (Rudy, 2009) and spatial planning information to other aspects of learning (Banquet et al., 2005; Hasselmo, 2005; Dollé et al., 2010; Martinet et al., 2011). It is also similar to points made by Redish and Touretzky (1998) that one can both store sequences and do location-recall in hippocampal attractor networks without interfering with each other (see also Redish, 1999).

Consequently, lesions of the hippocampus should affect both model-free and model-based systems through loss of spatial information, but transient interference with its activity should affect only the model-based system through loss of the use of the model. **Figure 4** illustrates how our proposition may account for the recent results obtained by Jadhav et al. (2012). In this study, rats experienced a W-track spatial alternation task: they alternated between "inbound" trials where they had to go to the center starting from either the left or the right arm and "outbound" trials where they had to go from the central arm to the arm (left or right) that they did not visit on the previous trial (**Figure 4A**). Outbound trials present a higher degree of difficulty in that they require linking past experience—the previously experienced side of the maze—with current location in order to make an appropriate decision. Strikingly, lesion of the hippocampus impaired both inbound and outbound learning (Kim and Frank, 2009) while disruption of awake hippocampal replay only impaired outbound learning (Jadhav et al., 2012).

We show on **Figure 4B** (resp. **C**) how a model-free (resp. model-based) system dependent on hippocampal input could explain the results. A model-free system learning the association between a spatial state (i.e., left arm, right arm, or central arm)

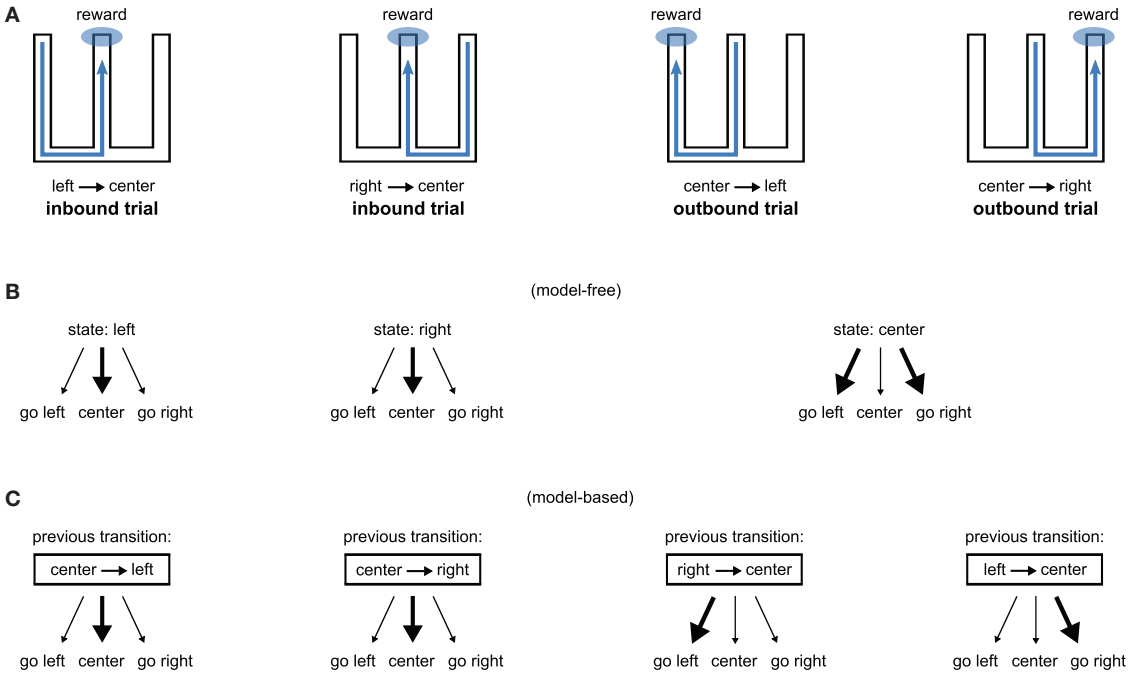


FIGURE 4 | Model-based/model-free framework applied to a spatial alternation task requiring both inbound and outbound learning. **(A)** W-shaped maze experienced by rats, adapted from Kim and Frank (2009). Hippocampal lesions impair both inbound and outbound learning (Kim and Frank, 2009) while disruption of awake hippocampal replay only impairs outbound learning (Jadhav et al., 2012). **(B)** A model-free system associating places with actions can learn inbound trials but would face high uncertainty during outbound trials. **(C)** A model-based system associating previous transitions with actions can associate past experience with current location and is thus able to learn both inbound and outbound trials.

and an action would be able to learn inbound trials but not outbound trials. This is because the "center" state is half of the time followed by rewarded trials on the left and half of the time followed by rewarded trials on the right, thus producing a situation with high uncertainty. In contrast, a model-based system learning to associate previous state transitions with actions can solve both inbound and outbound trials (**Figure 4C**). Thus, within our proposal, hippocampal lesions impair both inbound and outbound learning because they suppress spatial information required by both place-based model-free and model-based systems. By contrast, disruption of hippocampal awake replay would impair only the model-based system, potentially by blocking the storage of transitions in the model (Gupta et al., 2010), sparing the model-free system to still learn inbound trials.

5. VENTRAL STRIATUM—MODEL BUILDER?

What, then, might be the role of the VS in model-free and model-based navigation? Ventral striatal recordings and lesion studies have provided strong evidence for an evaluative role, either as part of the "critic" contributing to the calculation of the reward prediction error (O'Doherty et al., 2004; Khamassi et al., 2008), or as the locus for general Pavlovian-instrumental transfer where rewarded stimuli act to motivate future action (Corbit et al., 2001; Yin et al., 2008; Corbit and Balleine, 2011). The actor/critic architecture is a variant of the model-free reinforcement algorithms, which conceptually splits the value learning and action selection components (Sutton and Barto, 1998): the critic learns the value of every state, and uses those values to compute the reward prediction error after each state transition s to s', given any reward obtained; the prediction error is used by the actor to change the probability of selecting each action in state s, thus reflecting the outcome. The existing evidence that dorsal striatum supports action selection while the VS supports stimulus-outcome association has led to proposals that they respectively subserve the actor and critic roles (Joel et al., 2002; O'Doherty et al., 2004; Khamassi et al., 2005, 2008; Daw et al., 2011; van der Meer and Redish, 2011). The primary candidate for transmitting the reward prediction error is the phasic activity of the midbrain dopamine neurons (Schultz et al., 1997; Bayer and Glimcher, 2005; Cohen et al., 2012); further strengthening the proposed identification of the VS with the critic is that it is the major source of inputs to the dopamine neurons (Watabe-Uchida et al., 2012) that in turn project to the dorsal striatum (Maurin et al., 1999; Haber et al., 2000) (see **Figure 6**).

We sketch an account here that finesses this view, extending previous proposals (Yin et al., 2008; Bornstein and Daw, 2011) for separately considering the core and shell. We first argue that in addition to being useful for the "critic" in model-free processes, reward information encoded by the VS also contributes to model-based processes such as the building of a reward function. Second, from the perspective of navigation tasks, we find evidence that the core of the VS is a key locus for learning the correct sequences of actions in a task. A useful consequence of considering this proposed model-based/model-free dichotomy in both conditioning and navigation is that, whereas the core of the VS is often ascribed a purely evaluative role in the conditioning literature (Yin and Knowlton, 2006; Yin et al., 2008; Bornstein and Daw, 2011), the literature on core involvement in navigation clearly points to a major role in the direct control of locomotion. For the shell of the VS, we discuss further the suggestion that it is a key locus of the critic that signals the reward prediction error for the model-based system (Bornstein and Daw, 2011)[5]; we also discuss the possibility that it acts as a critic that signals a *state* prediction error in the predicted and actual state transitions. As these functions of the core and shell are essential for correct assemblage of the "model" of the world, we informally label the VS as part of the "model-builder".

5.1. VENTRAL STRIATUM AS SUBSTRATE FOR BUILDING THE REWARD FUNCTION

In the machine learning literature, one of the requirements for model-based algorithms is to build the so-called "reward function" which relates states to rewards [see **Figure 1**; (Sutton and Barto, 1998)]. In spatial tasks, this consists of memorizing the places in which reward is found. This is crucial information for deliberative decision-making where inference of future outcomes within the estimated world model—e.g., the tree-search process—requires reaching a terminal state where a reward can be found. The reward function is also important for off-line simulations within the world model to consolidate trajectories leading to reward—see for instance the *DynaQ* algorithm (Sutton and Barto, 1998). Indeed, such mental simulations should be informed when the agent has virtually reached a state containing a reward, although the agent is not necessarily physically experiencing such reward.

Interestingly, sequences of hippocampal place cell activations that occur while an animal is running a track in search for reward are known to be replayed during subsequent sleep (Euston et al., 2007) or during awake resting periods (Foster and Wilson, 2006; Gupta et al., 2010). These replay events have been hypothesized to participate in the consolidation of relevant behavioral sequences that lead to reward. Of particular interest for this review are recent reports of off-line synchronous replay between ventral striatal and hippocampal activity (Lansink et al., 2009). Lansink et al. (2009) found pairs of hippocampus—VS neurons that were reactivated during awake fast forward replay preferentially if: the hippocampal cell coded for space, the ventral striatal cell coded for reward, and the hippocampal cell was activated slightly before the ventral striatal cell during the task. The reactivation occurred 10 times faster than the sequence of activity during the task execution, possibly complying with physiologically plausible eligibility timing. The ventral striatal cells were predominantly in the core—but also included the shell. By illustrating possible neural mechanisms for the off-line consolidation of place-reward associations, these results provide striking examples of activity that could underly the building of the "reward function", which relates states to rewards.

[5]This relates to the notion, in the machine learning literature, that some model-based algorithms such as *Dyna-Q* can update their state-action values through a reward prediction error (RPE), although other model-based algorithms based on so-called *value iteration* processes do not rely on a RPE: they instead propagate value information from each state to other proximal states (Sutton and Barto, 1998).

Of course, it is plausible that such replay events could at the same time be used to update value estimations and action probabilities in the model-free system, consistent with the hypothesized *critic* role of part of the VS (O'Doherty et al., 2004; Khamassi et al., 2008; Bornstein and Daw, 2011). But if the ventral striatal part engaged during these replay events was only dedicated to model-free reinforcement learning, all ventral striatal cells encoding reward predictions in any location—not only in the reward location—should be reactivated in correspondence with the hippocampal cells coding for their associated states, which is not the case here. These results thus emphasize that the VS's evaluative role and its involvement in encoding reward information may also contribute to model-based processes. In support of this view, McDannald et al. (2011) recently showed in rats experiencing an unblocking procedure that VS not only incorporates information about reward value but also about specific features of the expected outcomes. Along with the orbitofrontal cortex, VS was indeed found to be required for learning driven by changes in reward identity, information only relevant for model-based processes but not for model-free ones which only work with value information.

Now where does the information which is replayed off-line between VS and hippocampus come from? One possibility is that relevant place-reward associations experienced during task performance are tagged in order to be preferentially replayed during subsequent sleep or awake resting periods. In support of this proposition, van der Meer and Redish (2010)'s synchronous recordings of VS and hippocampus in a T-maze disentangled possible mechanisms underlying the binding of hippocampal place representations and ventral striatal reward information during task performance. They found a ventral striatal phase precession relative to the hippocampal theta rhythm. This phase precession was found in ventral striatal ramp neurons preferentially receiving input from those hippocampal neurons that were active leading up to reward sites. This phenomenon was accompanied by increased theta coherence between VS and the hippocampus, possibly underlying the storage of relevant place-reward associations that should be tagged for subsequent consolidation.

5.2. VENTRAL STRIATAL CORE AS SUBSTRATE FOR BUILDING THE ACTION MODEL

Yin et al. (2008) proposed that one of the core's primary functions is to learn stimulus-outcome associations that drive preparatory behavior such as approach. Bornstein and Daw (2011) proposed in turn that, as preparatory behavior is value-agnostic, this is consistent with the core playing the role of the critic in a model-free controller: that it either computes directly or conveys the values of current and reached state to midbrain dopamine neurons (Joel et al., 2002), which in turn signal the reward prediction error to targets in the striatum and PFC (Schultz et al., 1997; Dayan and Niv, 2008). This proposal naturally extends to the core playing the role of model-free critic in navigation as well as conditioning.

However, it is equally clear that the core has a role in direct control of motor behavior, and may even serve as an action selection substrate separate from the dorsal striatum (see Pennartz et al., 1994; Nicola, 2007; Humphries and Prescott, 2010 for reviews). These dual roles for the core are not in conflict: the separate populations of core neurons that either project to the dopaminergic neurons of the midbrain or project to the other structures of the basal ganglia could, respectively, fulfill the evaluative and motor control roles (Humphries and Prescott, 2010). Here we focus on how the latter role may fit into a putative model-based/model-free separation of navigation based on the dorsal striatum.

It has long been known that core application of NMDA, AMPA, or dopamine agonists, or of drugs of abuse (amphetamine, cocaine), induces hyperlocomotion in rats, and that intact output of the core through the basal ganglia is necessary for this hyperlocomotion to occur (Pennartz et al., 1994; Humphries and Prescott, 2010). The phasic activity of individual core neurons also correlates with the onset of locomotion during self-administration of cocaine (Peoples et al., 1998). During behavioral tasks, the activity of individual neurons in the core correlates with the direction of upcoming movement, irrespective of the properties of the cue used to prompt that movement (Setlow et al., 2003; Taha et al., 2007). Moreover, when rats navigate a maze, the activity of core neurons correlates with the direction of movement in specific locations (Shibata et al., 2001; Mulder et al., 2004). Together, these data suggest that the core not only directly controls movement, but also receives spatial information on which to base that control.

In addition, the core is necessary for correctly learning sequences of motor behaviors. Blocking NMDA receptors in the core, which putatively prevents synaptic plasticity, degrades performance on many spatial tasks: rats cannot learn paths to rewards (Kelley, 1999), learn spatial sequences (in this case, of lever presses) to achieve reward (Bauter et al., 2003), or locate a hidden platform in a Morris water maze when encoded by distal cues alone (Sargolini et al., 2003). Lesioning hippocampal afferents to VS by cutting the fornix/fimbra pathway results in numerous spatial navigation problems. Whishaw and colleagues have shown that rats with such lesions have intact place responses, but great difficulty in constructing paths to them (Whishaw et al., 1995; Gorny et al., 2002). In a Morris water maze, lesioned rats can swim to a pre-lesion submerged platform location, but not to a new one (Whishaw et al., 1995); in open-field exploration, lesioned rats do not show path integration trips to their homebase (Gorny et al., 2002). Data from these studies has to be interpreted with care, but are consistent with the NMDA blockade studies. Together these data point to a key role for ventral striatal core in linking together sequential episodes of behavior.

So what is the motor control part of the core doing within the model-based/model-free framework? A general proposition is that the core is the route via which hippocampal sequencing of states reaches the motor system, a finessing of the long-recognized position of the core at the limbic-motor interface (Mogenson et al., 1980). We sketch a proposal here that its specific computational role is to learn and represent the probability of action selection within the transition model of the model-based system.

5.2.1. Actions in the transition model

Consider the transition model $T(s', a, s)$, giving the probability of arriving in state s' given action a and current state s; which we can also write $p(s'|a, s)$. The model has two uses: for

off-line learning, it is used to sample trajectories through the world model, and update the values of each state accordingly (Sutton and Barto, 1998; Johnson and Redish, 2005); for on-line action selection, it can be queried for the probability that each action will lead to the desired transition from state s to s'. To achieve this dual use it might be advantageous to decompose the transition model $p(s'|a, s)$ using Bayes theorem into representations of the state transitions and of the probability of action selection:

$$p(s'|a, s) = p(s'|s)\frac{p(a|s', s)}{p(a|s)},$$

where we assume that current state s is known. The first-term $p(s'|s)$ is then just the probability model for state transitions, the second term is just the probability $p(a|s', s)$ that each action will cause that transition, normalized by the probability $p(a|s)$ of ever taking that action in state s. Consequently, off-line learning is a product of the two terms, whereas on-line action selection can be based on the second term only.

Such a decomposition in turn suggests a decomposition into neural substrates. The hippocampal formation has long been proposed to represent potential state transitions (Poucet et al., 2004), and so is a natural candidate for representing $p(s'|s)$ in the simultaneous activity of current (s) and adjacent (s') place cells. Alternatively, neural network modeling of hippocampal formation functions in spatial navigation has even suggested that the directional-specificity of many place fields could be interpreted not as place cells but rather as "transition" cells, representing the possible transitions between the current and next "states" in the environment (Gaussier et al., 2002). In this account, each cell is a candidate for directly encoding $p(s'|s)$.

The ventral striatal core is then a potential substrate for representing the transition-conditioned probability of action selection $p(a|s', s)$. A plausible network implementation is that hippocampal outputs representing s and s' converge on neuron groups in the core, whose consequent activity is then proportional to $p(a|s', s)$. Learning this action component $p(a|s', s)$ of the transition model is then equivalent to changes in the synaptic weights linking the two state representations in hippocampus to the neuron group in the core. Over all known state transitions from the current state s, the activity in the core then encodes a probability distribution over potential actions; the selection of action based on this distribution is then done by the core's corresponding basal ganglia circuit (see Redgrave et al., 1999; Nicola, 2007; Humphries and Prescott, 2010; Humphries et al., 2012 for detailed models of this process).

This decomposition into substrates suggests that core neurons should thus show activity correlated with both off-line model search and on-line action selection. The latter we have already discussed: core activity is correlated with specific actions; in particular, the studies of Shibata et al. (2001) and Mulder et al. (2004) showing a set of core neurons with motor-related activity only in specific places within a maze (such as an arm), and then only when the rats move in a particular direction in that place (e.g., toward the arm end), are consistent with the encoding of action probability conditioned on a transition between states. This substrate decomposition also suggests that hippocampal formation and the core should be synchronized throughout free exploration, as continually changing states represented in hippocampus should have a corresponding recruitment of changing action selection probabilities in the core—just such an exploration-specific synchronization in local-field potentials between hippocampus and the core has been reported by Gruber et al. (2009). More electrophysiological studies will be required to confirm this hypothesis and precisely identify the underlying mechanisms.

Recent neurophysiological studies also support the existence of neural activity consistent with off-line model use for decision-making in the core. In a multiple T-maze, van der Meer and Redish (2009) found that neurons in the core which fired at either reward site also fired at the maze's decision point, just where hippocampal activity correlates of forward planning have been previously found (Johnson and Redish, 2007). Such activity at decision points occurred before reward was actually experienced, and thus before error correction. This activity appeared only during initial stages and disappeared after additional training producing behavioral automation. Such activity could thus reflect a search process related to the early use of model-based processes for decision-making by providing signals for the evaluation of internally generated possible transitions considered during navigation (van der Meer and Redish, 2009).

5.3. VENTRAL STRIATAL SHELL AS CRITIC(S) IN THE MODEL-BUILDER: ONE SYSTEM AMONGST MANY

More than any other region of the striatum, the ventral striatal shell is a complex intermingling of multiple separate systems (Humphries and Prescott, 2010), which may include control of approach and aversive behaviors (Reynolds and Berridge, 2003), hedonic information, outcome evaluation, memory consolidation, and appetitive control (Kelley, 1999). Consequently, we cannot meaningfully speak of *a* role for the shell; not least because, as we noted in Humphries and Prescott (2010), the lateral and medial shell are themselves easily distinguished entities in terms of their afferent and efferent structures—we will return to this distinction below.

Yin et al. (2008) proposed that the shell's primary function is to learn stimulus-outcome associations that drive consummatory behavior. Bornstein and Daw (2011) argued that this role in consummatory behavior requires a sensitivity to the values of the outcome, and thus makes the shell a natural candidate for subserving a role equivalent to the "critic" for the model-based system. While strictly speaking the actor/critic algorithm is a model-free system, the model-based system still may rely on the computation of a prediction error to update the values of each state (van der Meer and Redish, 2011), whether during off-line model search or on-line update after each performed action. Recently, Daw et al. (2011) tested human subjects on a multi-stage decision task that separated model-based and model-free prediction errors, and found that the model-based prediction error correlated with the fMRI BOLD signal in VS.

Against this idea, earlier work has shown that the shell appears not to be required for knowledge of the contingency between instrumental actions and their outcomes: lesioning the shell does

not stop devaluation or contingency changes from changing behavioral choice (Balleine and Killcross, 1994; Corbit et al., 2001). Consequently, the shell could appear not to be necessary for establishing goal-directed learning—or, by extension, model-based learning.

However, a closer reading of the lesion studies allows us to refine that conclusion. In "shell" lesion studies, only the medial shell is targeted (see, for example, Figure 1 of Corbit et al., 2001)—not a flaw in experimental design but a limitation imposed by anatomy, as attempts to lesion the lateral shell would undoubtedly also damage the overlying lateral core (Ikemoto, 2002). Consequently, the lateral shell remains intact, and is thus a prime candidate for a model-based critic that leaves the animal sensitive to outcome devaluation and contingency changes.

Moreover, as we detailed in Humphries and Prescott (2010), lateral and medial shell are separable entities: medial shell receives extensive input from hippocampal field CA1 and subiculum, while lateral shell receives scant hippocampal input; and both have separate "direct" and "indirect" pathways through the basal ganglia to separate populations of midbrain dopaminergic neurons (**Figure 5A**). As we show in **Figures 5B**,**C**, the dual pathways are a plausible candidate for computing a prediction error based on comparing the forebrain inputs to the two pathways; consequently both medial and lateral shell could support different "critic" roles (Humphries and Prescott, 2010).

Which leaves the question of the role of the medial shell, if it is indeed in a position to compute a prediction error. In Humphries and Prescott (2010) we proposed the idea that the projections from hippocampal formation and PFC to the "direct" and "indirect" pathways could, respectively, represent the expected and achieved state after a transition. Consequently, the medial shell would be in a position to compute a *state* prediction error, that adjusts the transition probability $p(s'|s)$ based on model predictions, rather than on simply counting the occurrences of each transition.

Lesioning the medial shell would then be predicted to show subtle deficits in tasks that require building a world model: in sufficiently simple tasks, the mere construction of the links between a limited number of states, whose values are correctly learnt, may be sufficient to solve the task and respond to subsequent changes in the value of those states. Consequently, the intact sensitivity to devaluation by medial shell-lesioned rats (Balleine and Killcross, 1994; Corbit et al., 2001) suggests that these were sufficiently simple tasks. That task complexity is a factor is suggested by the data of Albertin et al. (2000). They trained rats on a plus-maze on which a currently lit arm-end contained reward in the form of water drops; each day the rats experienced a new sequence of lit arms, and each day one of the arms was chosen to contain six drops and the others contained one drop. A probe trial was then run in which every arm was lit, allowing the rat to choose which arm to visit. Albertin et al. (2000) found that lesioning the medial shell prevented rats from correctly remembering which maze arm contained the high value reward on a probe trial, but did not impair their ability to learn to visit the lit arm in the sequence during training. Such a task plausibly requires each day building anew a world model and querying it on the probe trial to recall which available state-transition contained the high reward on that day. If damage to the medial shell prevented correct learning of the transition model, then this would selectively impair querying of the model, while leaving intact the

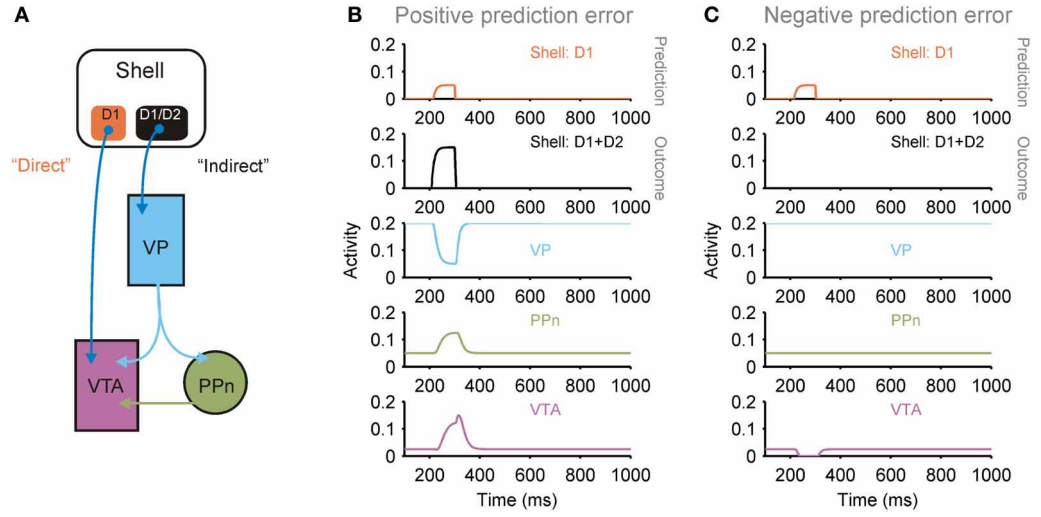


FIGURE 5 | Dual pathways from shell to ventral tegmental area (VTA) potentially support prediction error computation. (A) The medial and lateral shell both support a dual pathway circuit that converges on dopaminergic neurons in the VTA: a direct pathway originating from a population of D1 receptor expressing striatal projection neurons, and an indirect pathway originating from a mixed population of D1 and D2 receptor expressing striatal projection neurons [see (Humphries and Prescott, 2010) for review]. This arrangement is consistent with the shell's role as a "critic": the pathways support the computation of a prediction error between the prediction transmitted by the direct pathway and the actual outcome transmitted by the indirect pathway (PPn, pedunculopontine nucleus; VP, ventral pallidum). **(B)** Simulation of neural population activity showing how a greater outcome (indirect pathway) than predicted (direct pathway) drives a phasic increase in VTA activity, signaling a positive prediction error. **(C)** Simulation of neural population activity showing how a lower outcome (indirect pathway) than predicted (direct pathway) drives a phasic dip in VTA activity, signaling a negative prediction error. Simulation details given in Humphries and Prescott (2010).

ability to do simple light-reward association in the model-free system.

Glascher et al. (2010) searched for correlates of a state prediction error in the fMRI BOLD signal recorded from humans learning a decision-tree of stimulus choices in the absence of reward, which was subsequently used as the basis for a rewarded task. Encouragingly, subjects' behavior during the learning stage was well-fit by a reinforcement learning model incorporating a state prediction error; moreover, the BOLD signal in lateral PFC and intra-parietal sulcus correlated with the state prediction error in the model. The equivalent regions in rat are known afferents of the shell (Uylings et al., 2003; Humphries and Prescott, 2010). However, they reported that the ventral striatal BOLD signal correlated only with the fitted model-free reward prediction error during the rewarded task stage, and not the state prediction error. It is not clear, though, whether something computed by a set of neurons as small as the proposed sub-set in medial shell could be resolved by the voxel-size used, a problem compounded by the conservative multiple-comparison corrections used in searching for BOLD signal correlates.

6. CONCLUSIONS

In this paper, we have proposed a functional distinction between parts of the striatum by bridging data about their respective involvement in behavioral adaptation taken from both the spatial navigation literature and the instrumental conditioning literature. To do so, we have first formally mapped taxonomies of behavioral strategies from the two literatures to highlight that navigation strategies could be relevantly categorized as either model-based or model-free. At root, the key distinction is that it is the *use* of information in building a world representation, rather than the *type* of information (i.e., place vs. cue), that defines the different computational processes at stake and their substrates in the striatum. Within this framework, we explicitly identified the role for dorsolateral striatum in learning and expression of model-free strategies, the role of dorsomedial striatum in learning and expression of model-based strategies, and the role of "model-builder" for the VS—most probably in conjunction with the hippocampus (Lansink et al., 2009; van der Meer et al., 2010; Bornstein and Daw, 2012). Our scheme is summarized in **Figure 6**.

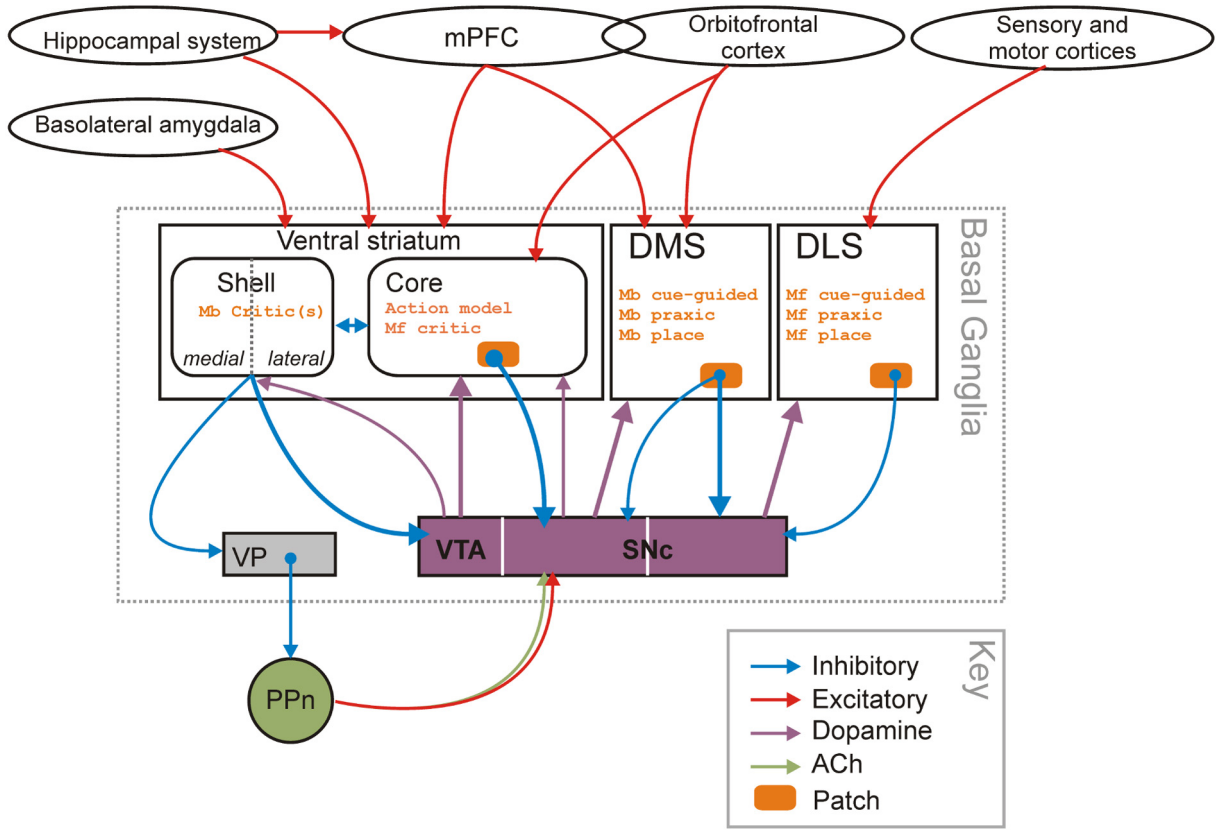


FIGURE 6 | Striatal-domain substrates of model-free and model-based controllers. The proposed organization of navigation strategies and potential control of learning across the three striatal domains. The identification of the shell and core as "critics" for the model-based and model-free controllers in dorsal striatum partly rests on the "spiral" of striatal-dopamine-striatal projections (Maurin et al., 1999; Haber et al., 2000; Haber, 2003), originating in the shell of the VS (the spiral is indicated by the thicker lines) and on the permissive role dopamine plays in plasticity at cortico-striatal synapses (Reynolds et al., 2001; Shen et al., 2008). There are also closed loop links between dopamine cell populations and each striatal region. Abbreviations: Mb, model-based; Mf, model-free; PPn, pedunculopontine nucleus; SNc, substantia nigra pars compacta; VP, ventral pallidum; VTA, ventral tegmental area. Note that the "inhibitory" and "excitatory" labels refer to the dominant neurotransmitter of the connection, not the effect that connection may have on the target nucleus as a whole (e.g., basolateral amygdala input to VS neurons can suppress other excitatory inputs despite using glutamate, which is an "excitatory" neurotransmitter).

The hypothesis that two decision-making systems (i.e., model-based and model-free) are processed in parallel in DMS and DLS while VS is important for the acquisition of the model seems to well explain the results of Atallah et al. (2007). In a forced-choice task in a Y-maze requiring rats to learn the association between two odors and two actions (go left or right), they found that transient inactivation of DLS[6] did not prevent a covert learning process which became visible as soon as the DLS was released. Although this task is typically interpreted as a habit learning task (van der Meer et al., 2012), the absence of over-training in the animals—60 trials performed in total—suggests that model-based learning in the DMS was still playing an important role at this stage and was unaffected by DLS inactivation. Moreover, Atallah et al. (2007) found that inactivation of VS mostly impaired acquisition and only partially affected performance, consistent with the proposed role of VS in building the model used by the model-based system.

6.1. COMPUTATIONS BY THE STRIATUM

Our proposed division of function between different parts of the striatum preserves the classical hypothesis that striatal territories all contribute to behavioral regulation but mainly differ in function because of their different afferents (Alexander et al., 1990; Joel and Weiner, 1994; Middleton and Strick, 2000)—a common division of cortical afferents among the striatal territories is illustrated in **Figure 6**. Throughout its dorso-lateral to ventro-medial extent, the striatum has a consistent micro-circuit dominated by GABAergic projection neurons controlled by at least three classes of interneurons (Tepper et al., 2004; Bolam et al., 2006; Humphries and Prescott, 2010). Such a consistent micro-architecture points to common operational principles for how striatum computes with its afferent inputs. Moreover, the cortex-basal ganglia-thalamus-cortex anatomical loop involving the ventral striatal core respects the same organization principles as loops involving the dorsal striatum: thus DLS, DMS, and VS core are all involved in complete basal ganglia circuits composed of direct and indirect pathways (Humphries and Prescott, 2010). Since numerous computational studies have shown that this basal ganglia circuitry is efficient for performing a selection process (Houk and Wise, 1995; Mink, 1996; Redgrave et al., 1999; Humphries et al., 2006; Leblois et al., 2006; Girard et al., 2008), it has been proposed that loops involving different striatal territories could perform different levels of selection influencing behavior. One such scheme envisions a hierarchy running from course-grained selection of overall goal or strategy to achieve a goal, through actions toward a goal, to fine-grained movement parameters of each action (Redgrave et al., 1999; Ito and Doya, 2011).

[6]Although the injection site was referred to as the central part of the dorsal striatum by the authors (see Supplementary Figures 3 and 4 of their original paper), the great majority of injections were located outside the dorsal striatal region receiving projections from the prelimbic cortex [see Figure 3 in Voorn et al. (2004)], and thus outside the zone called dorsomedial striatum and related to goal-directed behaviors and model-based learning [see Figure 1 in Yin et al. (2008) and Figure 1 in Bornstein and Daw (2011)]. Thus, the injections seem to have mostly reached the dorsolateral striatum related to model-free habit learning.

The model-based/model-free dichotomy would respect such a general principle of common selection operation: that striatal territories receiving state transition information (i.e., $p(s'|s)$ corresponding to the probability of transition from state s to state s', no matter if these states are spatial or determined by a perceptual cue) would be involved in model-based action selection while striatal territories receiving simple state information (i.e., $p(s)$, no matter if state s represents a spatial position or the perception of a stimulus) would be involved in model-free action selection. As we discussed throughout the text, in contrast to DLS, VS and DMS receive direct projections from the hippocampal system as well as medial PFC which place them in a good situation to process hippocampal state transition information (Gaussier et al., 2002; Poucet et al., 2004) and hence to participate in the model-based action selection. Correspondingly, the dominant projections of sensorimotor cortices to DLS may thus convey current state information, whether originating from the periphery or from higher cortical areas (Haber, 2003), and hence the DLS participates in model-free action selection.

6.2. OPEN QUESTIONS

The account here provides concrete proposals for the dorsolateral and dorsomedial striatum's role in spatial navigation, while introducing new but comparatively speculative ideas about the VS's roles in the model-free and model-based systems. As such, our account is of course incomplete; so let us conclude with the primary open questions:

- We have drawn a distinction between place/response strategies and model-based/model-free use of those strategies. To the best of our knowledge, we lack good evidence for the existence of a model-free place strategy.
- The observations of a place-to-response strategy shift with over-training (Dickinson, 1980; Packard and McGaugh, 1996; Pearce et al., 1998; Chang and Gold, 2003) underpinned the existing idea that a response strategy is by nature habitual. Our hypothesis postulates that the central mechanism underlying all these observed behavioral shifts is a shift from model-based to model-free rather than from place-based to either cue-guided or praxic behaviors; but why then is the shift often (but not always Yin and Knowlton, 2004; Botreau and Gisquet-Verrier, 2010) from model-based place to model-free response?
- What is anterior DMS doing? Ragozzino and Choi (2004) proposed a role for it in strategy selection, as lesions caused a selective deficit in reversal learning, but not in initial acquisition. Alternatively, perhaps DMS is divided into sub-territories differentially involved in place, cue, and praxic model-based systems.
- Lesion data on the core provide conflicting accounts of its roles. For example, the results of Corbit et al. (2001) disagree with evaluation: for why, if the core forms part of the transition model, does lesioning it not then prevent outcome devaluation from affecting behavior? By contrast, McDannald et al. (2011) found that lesions of core affected responding to both changes in outcome value and changes in outcome identity, emphasizing its involvement in model-based learning.

From our account, it is not surprising that conflicting data arise if core lesions interfere with both evaluative and action selection systems; however, it is not clear what task designs would be sufficient to tease apart the selective effects of core lesions on its evaluative and action selection roles.

- Do the striatal domains underpin a common computation? Our focus has been on the algorithmic-level distinctions between behavioral strategies, and the striatal substrates within the neural systems implementing those algorithms. As noted throughout, this computation may be action selection: the resolution of competing inputs at the striatal level into one (or a few) selected signals at the output of the basal ganglia. Based on our proposals here, we may speculate that these selections are based on different representations of the world.

ACKNOWLEDGMENTS

This work was supported by L'Agence Nationale de la Recherche: ANR-11-BSV4-006 "LU2" (Learning Under Uncertainty) project (Mehdi Khamassi), and ANR-2010-BLAN-0217-04 "NEUROBOT" project (Mark D. Humphries); by the "HABOT" project of the Ville de Paris Emergence(s) program (Mehdi Khamassi); by a MRC Senior non-Clinical Fellowship (Mark D. Humphries); and by the European Community FP6 IST 027819 "ICEA" (Integrating Cognition Emotion and Autonomy) Project (Mark D. Humphries and Mehdi Khamassi).

REFERENCES

Adams, S., Kesner, R. P., and Ragozzino, M. E. (2001). Role of the medial and lateral caudate-putamen in mediating an auditory conditional response association. *Neurobiol. Learn. Mem.* 76, 106–116.

Albertin, S. V., Mulder, A. B., Tabuchi, E., Zugaro, M. B., and Wiener, S. I. (2000). Lesions of the medial shell of the nucleus accumbens impair rats in finding larger rewards, but spare reward-seeking behavior. *Behav. Brain Res.* 117, 173–183.

Alexander, G. E., Crutcher, M. D., and DeLong, M. R. (1990). Basal ganglia-thalamocortical circuits: parallel substrates for motor, oculomotor, "prefrontal" and "limbic" functions. *Prog. Brain Res.* 85, 119–146.

Arleo, A., and Gerstner, W. (2000). Spatial cognition and neuromimetic navigation: a model of hippocampal place cell activity. *Biol. Cybern.* 83, 287–299.

Arleo, A., and Rondi-Reig, L. (2007). Multimodal sensory integration and concurrent navigation strategies for spatial cognition in real and artificial organisms. *J. Int. Neurosci.* 6, 327–366.

Atallah, H. E., Lopez-Paniagua, D., Rudy, J. W., and O'Reilly, R. C. (2007). Separate neural substrates for skill learning and performance in the ventral and dorsal striatum. *Nat. Neurosci.* 10, 126–131.

Balleine, B. (2005). Neural bases of food-seeking: affect, arousal and reward in corticostriatolimbic circuits. *Physiol. Behav.* 86, 717–730.

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

Balleine, B., and Killcross, S. (1994). Effects of ibotenic acid lesions of the nucleus accumbens on instrumental action. *Behav. Brain Res.* 65, 181–193.

Banquet, J. P., Gaussier, P., Quoy, M., Revel, A., and Burnod, Y. (2005). A hierarchy of associations in hippocampo-cortical systems: cognitive maps and navigation strategies. *Neural Comput.* 17, 1339–1384.

Barnes, T. D., Kubota, Y., Hu, D., Jin, D. Z., and Graybiel, A. M. (2005). Activity of striatal neurons reflects dynamic encoding and recoding of procedural memories. *Nature* 437, 1158–1161.

Bauter, M. R., Brockel, B. J., Pankevich, D. E., Virgolini, M. B., and Cory-Slechta, D. A. (2003). Glutamate and dopamine in nucleus accumbens core and shell: sequence learning versus performance. *Neurotoxicology* 24, 227–243.

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

Berke, J. D., Breck, J. T., and Eichenbaum, H. (2009). Striatal versus hippocampal representations during win-stay maze performance. *J. Neurophysiol.* 101, 1575–1587.

Bolam, J. P., Bergman, H., Graybiel, A. M., Kimura, M., Plenz, D., Seung, H. S., et al. (2006). "Microcircuits in the striatum," in *Microcircuits: The Interface Between Neurons and Global Brain Function*, eds S. Grillner and A. M. Graybiel (Cambridge, MA: MIT Press), 165–190.

Bornstein, A. M., and Daw, N. D. (2011). Multiplicity of control in the basal ganglia: computational roles of striatal subregions. *Curr. Opin. Neurobiol.* 21, 374–380.

Bornstein, A. M., and Daw, N. D. (2012). Dissociating hippocampal and striatal contributions to sequential prediction learning. *Eur. J. Neurosci.* 35, 1011–1023.

Botreau, F., and Gisquet-Verrier, P. (2010). Re-thinking the role of the dorsal striatum in egocentric/response strategy. *Front. Behav. Neurosci.* 4:7. doi: 10.3389/neuro.08.007.2010

Brown, L., and Sharp, F. (1995). Metabolic mapping of rat striatum: somatotopic organization of sensorimotor activity. *Brain Res.* 686, 207–222.

Burgess, N., Recce, M., and O'Keefe, J. (1994). A model of hippocampal function. *Neural Netw.* 7, 1065–1081.

Chang, Q., and Gold, P. E. (2003). Switching memory systems during learning: changes in patterns of brain acetylcholine release in the hippocampus and striatum in rats. *J. Neurosci.* 23, 3001–3005.

Chang, Q., and Gold, P. E. (2004). Inactivation of dorsolateral striatum impairs acquisition of response learning in cue-deficient, but not cue-available, conditions. *Behav. Neurosci.* 118, 383–388.

Cohen, J. Y., Haesler, S., Vong, L., Lowell, B. B., and Uchida, N. (2012). Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature* 482, 85–88.

Corbit, L. H., and Balleine, B. W. (2011). The general and outcome-specific forms of pavlovian-instrumental transfer are differentially mediated by the nucleus accumbens core and shell. *J. Neurosci.* 31, 11786–11794.

Corbit, L. H., Muir, J. L., and Balleine, B. W. (2001). The role of the nucleus accumbens in instrumental conditioning: evidence of a functional dissociation between accumbens core and shell. *J. Neurosci.* 21, 3251–3260.

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

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

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

Dayan, P., and Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. *Curr. Opin. Neurobiol.* 18, 185–196.

De Leonibus, E., Costantini, V. J. A., Massaro, A., Mandolesi, G., Vanni, V., Luvisetto, S., et al. (2011). Cognitive and neural determinants of response strategy in the dual-solution plus-maze task. *Learn. Mem.* 18, 241–244.

De Leonibus, E., Oliverio, A., and Mele, A. (2005). A study on the role of the dorsal striatum and the nucleus accumbens in allocentric and egocentric spatial memory consolidation. *Learn. Mem.* 12, 491–503.

Devan, B. D., and White, N. M. (1999). Parallel information processing in the dorsal striatum: relation to hippocampal function. *J. Neurosci.* 19, 2789–2798.

Dickinson, A. (1980). *Contemporary Animal Learning Theory.* Cambridge, UK: Cambridge University Press.

Dickinson, A. (1985). Actions and habits: the development of behavioural autonomy. *Philos. Trans. R. Soc. B Biol. Sci.* 308, 67–78.

Dollé, L., Sheynikhovich, D., Girard, B., Chavarriaga, R., and Guillot, A. (2010). Path planning versus cue responding: a bio-inspired model of switching between navigation strategies. *Biol. Cybern.* 103, 299–317.

Euston, D. R., Tatsuno, M., and McNaughton, B. L. (2007).

Fast-forward playback of recent memory sequences in prefrontal cortex during sleep. *Science* 318, 1147–1150.

Faure, A., Haberland, U., Condé, F., and Massioui, N. E. (2005). Lesion to the nigrostriatal dopamine system disrupts stimulus-response habit formation. *J. Neurosci.* 25, 2771–2780.

Foster, D., Morris, R., and Dayan, P. (2000). Models of hippocampally dependent navigation using the temporal difference learning rule. *Hippocampus* 10, 1–16.

Foster, D. J., and Wilson, M. A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature* 440, 680–683.

Franz, M. O., and Mallot, H. A. (2000). Biomimetic robot navigation. *Rob. Auton. Syst.* 30, 133–153.

Gallistel, C. R. (1990). *The Organization of Learning*. Cambridge, MA: MIT Press.

Gaussier, P., Revel, A., Banquet, J. P., and Babeau, V. (2002). From view cells and place cells to cognitive map learning: processing stages of the hippocampal system. *Biol. Cybern.* 86, 15–28.

Girard, B., Tabareau, N., Pham, Q. C., Berthoz, A., and Slotine, J. J. (2008). Where neuroscience and dynamic system theory meet autonomous robotics: a contracting basal ganglia model for action selection. *Neural Netw.* 21, 628–641.

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

Gorny, J. H., Gorny, B., Wallace, D. G., and Whishaw, I. Q. (2002). Fimbria-fornix lesions disrupt the dead reckoning (homing) component of exploratory behavior in mice. *Learn. Mem.* 9, 387–394.

Graybiel, A. M. (1998). The basal ganglia and chunking of action repertoires. *Neurobiol. Learn. Mem.* 70, 119–136.

Groenewegen, H. J., Wright, C. I., and Beijer, A. V. (1996). The nucleus accumbens: gateway for limbic structures to reach the motor system? *Prog. Brain Res.* 107, 485–511.

Gruber, A. J., Hussain, R. J., and O'Donnell, P. (2009). The nucleus accumbens: a switchboard for goal-directed behaviors. *PLoS ONE* 4:e5062. doi: 10.1371/journal.pone.0005062

Gupta, A. S., van der Meer, M. A., Touretzky, D. S., and Redish, A. D. (2010). Hippocampal replay is not a simple function of experience. *Neuron* 65, 695–705.

Haber, S. N. (2003). The primate basal ganglia: parallel and integrative networks. *J. Chem. Neuroanat.* 26, 317–330.

Haber, S. N., Fudge, J. L., and McFarland, N. R. (2000). Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *J. Neurosci.* 20, 2369–2382.

Hannesson, D. K., and Skelton, R. W. (1998). Recovery of spatial performance in the morris water maze following bilateral transection of the fimbria/fornix in rats. *Behav. Brain Res.* 90, 35–56.

Hartley, T., and Burgess, N. (2005). Complementary memory systems: competition, cooperation and compensation. *Trends Neurosci.* 28, 169–170.

Hasselmo, M. (2005). A model of prefrontal cortical mechanisms for goal-directed behavior. *J. Cogn. Neurosci.* 17, 1115–1129.

Heimer, L., Alheid, G. F., de Olmos, J. S., Groenewegen, H., Haber, S., E., Harlan, R. E., et al. (1997). The accumbens: beyond the core-shell dichotomy. *J. Neuropsychiatry Clin. Neurosci.* 9, 354–381.

Hok, V., Save, E., Lenck-Santini, P. P., and Poucet, B. (2005). Coding for spatial goals in the prelimbic/infralimbic area of the rat frontal cortex. *PNAS* 102, 4602–4607.

Honzik, C. H. (1936). The sensory basis of maze learning in rats. *Comp. Psychol. Monogr.* 13, 113.

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

Humphries, M. D., Khamassi, M., and Gurney, K. (2012). Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. *Front. Neurosci.* 6:9. doi: 10.3389/fnins.2012.00009

Humphries, M. D., and Prescott, T. J. (2010). The ventral basal ganglia, a selection mechanism at the crossroads of space, strategy, and reward. *Prog. Neurobiol.* 90, 385–417.

Humphries, M. D., Stewart, R. D., and Gurney, K. N. (2006). A physiologically plausible model of action selection and oscillatory activity in the basal ganglia. *J. Neurosci.* 26, 12921–12942.

Ikemoto, S. (2002). Ventral striatal anatomy of locomotor activity induced by cocaine, (d)-amphetamine, dopamine and d1/d2 agonists. *Neuroscience* 113, 939–955.

Ito, M., and Doya, K. (2011). Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit. *Curr. Opin. Neurobiol.* 21, 368–373.

Jacobson, T. K., Gruenbaum, B. F., and Markus, E. J. (2012). Extensive training and hippocampus or striatum lesions: effect on place and response strategies. *Physiol. Behav.* 105, 645–652.

Jadhav, S. P., Kemere, C., German, P. W., and Frank, L. M. (2012). Awake hippocampal sharp-wave ripples support spatial memory. *Science* 336, 1454–1458.

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

Joel, D., and Weiner, I. (1994). The organization of the basal ganglia-thalamocortical circuits: open interconnected rather than closed segregated. *Neuroscience* 63, 363–379.

Joel, D., and Weiner, I. (2000). The connections of the dopaminergic system with the striatum in rats and primates: an analysis with respect to the functional and compartmental organization of the striatum. *Neuroscience* 96, 451–474.

Jog, M. S., Kubota, Y., Connolly, C. I., Hillegaart, V., and Graybiel, A. M. (1999). Building neural representations of habits. *Science* 286, 1745–1749.

Johnson, A., and Redish, A. D. (2005). Hippocampal replay contributes to within session learning in a temporal difference reinforcement learning model. *Neural Netw.* 18, 1163–1171.

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

Kelley, A. E. (1999). Neural integrative activities of nucleus accumbens subregions in relation to learning and motivation. *Psychobiology* 27, 198–213.

Khamassi, M. (2007). *Complementary Roles of the Rat Prefrontal Cortex and Striatum in Reward-based Learning and Shifting Navigation Strategies*. PhD thesis, Université Pierre et Marie Curie.

Khamassi, M., Lacheze, L., Girard, B., Berthoz, A., and Guillot, A. (2005). Actor-critic models of reinforcement learning in the basal ganglia: from natural to arificial rats. *Adapt. Behav.* 13, 131–148.

Khamassi, M., Mulder, A. B., Tabuchi, E., Douchamps, V., and Wiener, S. I. (2008). Anticipatory reward signals in ventral striatal neurons of behaving rats. *Eur. J. Neurosci.* 28, 1849–1866.

Kim, S. M., and Frank, L. M. (2009). Hippocampal lesions impair rapid learning of a continuous spatial alternation task. *PLoS ONE* 4:e5494. doi: 10.1371/journal.pone.0005494

Kimchi, E. Y., and Laubach, M. (2009). Dynamic encoding of action selection by the medial striatum. *J. Neurosci.* 29, 3148–3159.

Kimchi, E. Y., Torregrossa, M. M., Taylor, J. R., and Laubach, M. (2009). Neuronal correlates of instrumental learning in the dorsal striatum. *J. Neurophysiol.* 102, 475–489.

Krech, D. (1932). The genesis of "hypotheses" in rats. *Publ. Psychol.* 6, 45–64.

Lansink, C. S., Goltstein, P. M., Lankelma, J. V., McNaughton, B. L., and Pennartz, C. M. A. (2009). Hippocampus leads ventral striatum in replay of place-reward information. *PLoS Biol.* 7:e1000173. doi: 10.1371/journal.pbio.1000173

Leblois, A., Boraud, T., Meissner, W., Bergman, H., and Hansel, D. (2006). Competition between feedback loops underlies normal and pathological dynamics in the basal ganglia. *J. Neurosci.* 26, 3567–3583.

Lex, B., Sommer, S., and Hauber, W. (2011). The role of dopamine in the dorsomedial striatum in place and response learning. *Neuroscience* 172, 212–218.

Martel, G., Blanchard, J., Mons, N., Gastambide, F., Micheau, J., and Guillou, J. (2007). Dynamic interplays between memory systems depend on practice: the hippocampus is not always the first to provide solution. *Neuroscience* 150, 743–753.

Martinet, L.-E., Sheynikhovich, D., Benchenane, K., and Arleo, A. (2011). Spatial learning and action planning in a prefrontal cortical network model. *PLoS Comput. Biol.* 7:e1002045. doi: 10.1371/journal.pcbi.1002045

Maurin, Y., Banrezes, B., Menetrey, A., Mailly, P., and Deniau, J. M. (1999). Three-dimensional distribution of nigrostriatal neurons in the rat: relation to the topography of striatonigral projections. *Neuroscience* 91, 891–909.

McDannald, M. A., Lucantonio, F., Burke, K. A., Niv, Y., and Schoenbaum, G. (2011). Ventral striatum and orbitofrontal cortex are both required for model-based,

but not model-free, reinforcement learning. *J. Neurosci.* 31, 2700–2705.
Middleton, F. A., and Strick, P. L. (2000). Basal ganglia and cerebellar loops: motor and cognitive circuits. *Brain Res. Brain Res. Rev.* 31, 236–250.
Mink, J. W. (1996). The basal ganglia: focused selection and inhibition of competing motor programs. *Prog. Neurobiol.* 50, 381–425.
Mogenson, G. J., Jones, D. L., and Yim, C. Y. (1980). From motivation to action: functional interface between the limbic system and the motor system. *Prog. Neurobiol.* 14, 69–97.
Morris, R. G. M. (1981). Spatial localization does not require the presence of local cues. *Learn. Motiv.* 12, 239–260.
Moussa, R., Poucet, B., Amalric, M., and Sargolini, F. (2011). Contributions of dorsal striatal subregions to spatial alternation behavior. *Learn. Mem.* 18, 444–451.
Mulder, A. B., Tabuchi, E., and Wiener, S. I. (2004). Neurons in hippocampal afferent zones of rat striatum parse routes into multi-pace segments during maze navigation. *Eur. J. Neurosci.* 19, 1923–1932.
Nicola, S. M. (2007). The nucleus accumbens as part of a basal ganglia action selection circuit. *Psychopharmacology (Berl.)* 191, 521–550.
O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* 304, 452–454.
O'Keefe, J., and Nadel, L. (1978). *The Hippocampus as a Cognitive Map*. Oxford, UK: Oxford University Press.
Packard, M. (1999). Glutamate infused posttraining into the hippocampus or caudate-putamen differentially strengthens place and response learning. *PNAS* 96, 12881–12886.
Packard, M., and McGaugh, J. (1992). Double dissociation of fornix and caudate nucleus lesions on acquisition of two water maze tasks: further evidence for multiple memory systems. *Behav. Neurosci.* 106, 439–446.
Packard, M., and McGaugh, J. (1996). Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects the expression of place and response learning. *Neurobiol. Learn. Mem.* 65, 65–72.
Packard, M. G., Hirsh, R., and White, N. M. (1989). Differential effects of fornix and caudate nucleus lesions on two radial maze tasks: evidence for multiple memory systems. *J. Neurosci.* 9, 1465–1472.
Packard, M. G., and Knowlton, B. J. (2002). Learning and memory functions of the basal ganglia. *Annu. Rev. Neurosci.* 25, 563–593.
Pearce, J. M., Roberts, A. D., and Good, M. (1998). Hippocampal lesions disrupt navigation based on cognitive maps but not heading vectors. *Nature* 396, 75–77.
Pennartz, C. M., Groenewegen, H. J., and da Silva, F. H. L. (1994). The nucleus accumbens as a complex of functionally distinct neuronal ensembles: an integration of behavioural, electrophysiological and anatomical data. *Prog. Neurobiol.* 42, 719–761.
Penner, M. R., and Mizumori, S. J. Y. (2012). Neural systems analysis of decision making during goal-directed navigation. *Prog. Neurobiol.* 96, 96–135.
Peoples, L. L., Gee, F., Bibi, R., and West, M. O. (1998). Phasic firing time locked to cocaine self-infusion and locomotion: dissociable firing patterns of single nucleus accumbens neurons in the rat. *J. Neurosci.* 18, 7588–7598.
Ploeger, G. E., Spruijt, B. M., and Cools, A. R. (1994). Spatial localization in the morris water maze in rats: acquisition is affected by intra-accumbens injections of the dopaminergic antagonist haloperidol. *Behav. Neurosci.* 108, 927–934.
Potegal, M. (1972). The caudate nucleus egocentric localization system. *Acta Neurobiol. Exp.* 32, 479–494.
Poucet, B., Lenck-Santini, P. P., Hok, V., Save, E., Banquet, J. P., Gaussier, P., et al. (2004). Spatial navigation and hippocampal place cell firing: the problem of goal encoding. *Rev. Neurosci.* 15, 89–107.
Pych, J. C., Chang, Q., Colon-Rivera, C., and Gold, P. E. (2005). Acetylcholine release in hippocampus and striatum during testing on a rewarded spontaneous alternation task. *Neurobiol. Learn. Mem.* 84, 93–101.
Ragozzino, M. E., and Choi, D. (2004). Dynamic changes in acetylcholine output in the medial striatum during place reversal learning. *Learn. Mem.* 11, 70–77.
Redgrave, P., Prescott, T. J., and Gurney, K. (1999). The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience* 89, 1009–1023.
Redish, A. D. (1999). *Beyond the Cognitive Map: From Place Cells to Episodic Memory*. Cambridge, MA: MIT Press.
Redish, A. D., and Touretzky, D. S. (1997). Cognitive maps beyond the hippocampus. *Hippocampus* 7, 15–35.
Redish, A. D., and Touretzky, D. S. (1998). The role of the hippocampus in solving the morris water maze. *Neural Comput.* 10, 73–111.
Reynolds, J. N., Hyland, B. I., and Wickens, J. R. (1957). Discrimination of cues in mazes: a resolution of the "place-vs.-response" question. *Psychol. Rev.* 64, 217–228.
Reynolds, J. N., Hyland, B. I., and Wickens, J. R. (2001). A cellular mechanism of reward-related learning. *Nature* 413, 67–70.
Reynolds, S. M., and Berridge, K. C. (2003). Glutamate motivational ensembles in nucleus accumbens: rostrocaudal shell gradients of fear and feeding. *Eur. J. Neurosci.* 17, 2187–2200.
Rudy, J. W. (2009). Context representations, context functions, and the parahippocampal-hippocampal system. *Learn. Mem.* 16, 573–585.
Sargolini, F., Florian, C., Oliverio, A., Mele, A., and Roullet, P. (2003). Differential involvement of NMDA and AMPA receptors within the nucleus accumbens in consolidation of information necessary for place navigation and guidance strategy of mice. *Learn. Mem.* 10, 285–292.
Schmitzer-Torbert, N. C., and Redish, A. D. (2008). Task-dependent encoding of space and events by striatal neurons is dependent on neural subtype. *Neuroscience* 153, 349–360.
Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science* 275, 1593–1599.
Setlow, B., and McGaugh, J. (1998). Sulpiride infused into the nucleus accumbens posttraining impairs memory of spatial water maze training. *Behav. Neurosci.* 112, 603–610.
Setlow, B., Schoenbaum, G., and Gallagher, M. (2003). Neural encoding in ventral striatum during olfactory discrimination learning. *Neuron* 38, 625–636.
Shen, W., Flajolet, M., Greengard, P., and Surmeier, D. J. (2008). Dichotomous dopaminergic control of striatal synaptic plasticity. *Science* 321, 848–851.
Shibata, R., Mulder, A. B., Trullier, O., and Wiener, S. I. (2001). Position sensitivity in phasically discharging nucleus accumbens neurons of rats alternating between tasks requiring complementary types of spatial cues. *Neuroscience* 108, 391–411.
Smith, D. M., and Mizumori, S. J. Y. (2006). Hippocampal place cells, context, and episodic memory. *Hippocampus* 16, 716–729.
Sutherland, R. J., and Hamilton, D. A. (2004). Rodent spatial navigation: at the crossroads of cognition and movement. *Neurosci. Biobehav. Rev.* 28, 687–697.
Sutherland, R. J., and Rodriguez, A. J. (1989). The role of the fornix/fimbria and some related subcortical structures in place learning and memory. *Behav. Brain Res.* 32, 265–277.
Sutton, R. S., and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.
Taha, S. A., Nicola, S. M., and Fields, H. L. (2007). Cue-evoked encoding of movement planning and execution in the rat nucleus accumbens. *J. Physiol.* 584, 801–818.
Tang, C., Pawlak, A. P., Prokopenko, V., and West, M. O. (2007). Changes in activity of the striatum during formation of a motor habit. *Eur. J. Neurosci.* 25, 1212–1227.
Tepper, J. M., Koos, T., and Wilson, C. J. (2004). GABAergic microcircuits in the neostriatum. *Trends Neurosci.* 27, 662–669.
Thorn, C. A., Atallah, H., Howe, M., and Graybiel, A. M. (2010). Differential dynamics of activity changes in dorsolateral and dorsomedial striatal loops during learning. *Neuron* 66, 781–795.
Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychol. Rev.* 55, 189–208.
Trullier, O., Wiener, S., Berthoz, A., and Meyer, J.-A. (1997). Biologically-based artificial navigation systems: review and prospects. *Prog. Neurobiol.* 51, 483–544.
Uylings, H. B. M., Groenewegen, H. J., and Kolb, B. (2003). Do rats have a prefrontal cortex? *Behav. Brain Res.* 146, 3–17.
van der Meer, M. A. A., Johnson, A., Schmitzer-Torbert, N. C., and Redish, A. D. (2010). Triple dissociation of information processing in dorsal striatum, ventral striatum, and hippocampus on a learned spatial decision task. *Neuron* 67, 25–32.
van der Meer, M. A. A., Kurth-Nelson, Z., and Redish, A. D. (2012). Information processing in decision-making systems. *Neuroscientist* 18, 342–359.
van der Meer, M. A. A., and Redish, A. D. (2009). Covert expectation-of-reward in rat ventral striatum at decision points. *Front. Integr. Neurosci.* 3:1. doi: 10.3389/neuro.07.001.2009

van der Meer, M. A. A., and Redish, A. D. (2010). Theta phase precession in rat ventral striatum links place and reward information. *J. Neurosci.* 31, 2843–2854.

van der Meer, M. A. A., and Redish, A. D. (2011). Ventral striatum: a critical look at models of learning and evaluation. *Curr. Opin. Neurobiol.* 21, 387–392.

Voorn, P., Vanderschuren, L. J., Groenewegen, H. J., Robbins, T. W., and Pennartz, C. M. (2004). Putting a spin on the dorsal-ventral divide of the striatum. *Trends Neurosci.* 27, 468–474.

Watabe-Uchida, M., Zhu, L., Ogawa, S. K., Vamanrao, A., and Uchida, N. (2012). Whole-brain mapping of direct inputs to midbrain dopamine neurons. *Neuron* 74, 858–873.

Whishaw, I. Q., Cassel, J. C., and Jarrad, L. E. (1995). Rats with fimbria-fornix lesions display a place response in a swimming pool: a dissociation between getting there and knowing where. *J. Neurosci.* 15, 5779–5788.

Whishaw, I. Q., Mittleman, G., Bunch, S. T., and Dunnett, S. B. (1987). Impairments in the acquisition, retention and selection of spatial navigation strategies after medial caudate-putamen lesions in rats. *Behav. Brain Res.* 24, 125–138.

White, N. M., and McDonald, R. J. (2002). Multiple parallel memory systems in the brain of the rat. *Neurobiol. Learn. Mem.* 77, 125–184.

Wiener, S. I. (1993). Spatial and behavioral correlates of striatal neurons in rats performing a self-initiated navigation task. *J. Neurosci.* 13, 3802–3817.

Wiener, S. I., Paul, C. A., and Eichenbaum, H. (1989). Spatial and behavioral correlates of hippocampal neuronal activity. *J. Neurosci.* 9, 2737–2763.

Willingham, D. B. (1998). What differentiates declarative and procedural memories: reply to cohen, poldrack, and eichenbaum (1997). *Memory* 6, 689–699.

Yin, H. H., and Knowlton, B. J. (2004). Contributions of striatal subregions to place and response learning. *Learn. Mem.* 11, 459–463.

Yin, H. H., and Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nat. Rev. Neurosci.* 7, 464–476.

Yin, H. H., Knowlton, B. J., and Balleine, B. W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *Eur. J. Neurosci.* 19, 181–189.

Yin, H. H., Knowlton, B. J., and Balleine, B. W. (2005a). Blockade of NMDA receptors in the dorsomedial striatum prevents action-outcome learning in instrumental conditioning. *Eur. J. Neurosci.* 22, 505–512.

Yin, H. H., Ostlund, S. B., Knowlton, B. J., and Balleine, B. W. (2005b). The role of the dorsomedial striatum in instrumental conditioning. *Eur. J. Neurosci.* 22, 513–523.

Yin, H. H., Ostlund, S. B., and Balleine, B. W. (2008). Reward-guided learning beyond dopamine in the nucleus accumbens: the integrative functions of cortico-basal ganglia networks. *Eur. J. Neurosci.* 28, 1437–1448.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 15 May 2012; accepted: 29 October 2012; published online: 27 November 2012.

Citation: Khamassi M and Humphries MD (2012) Integrating cortico-limbic-basal ganglia architectures for learning model-based and model-free navigation strategies. Front. Behav. Neurosci. ***6****:79. doi: 10.3389/fnbeh.2012.00079*

2.2 Parallel learning during Pavlovian conditioning

2.2.1 Lesaint, Sigaud, Flagel, Robinson, Khamassi (2014) PLoS Computational Biology

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Modelling Individual Differences in the Form of Pavlovian Conditioned Approach Responses: A Dual Learning Systems Approach with Factored Representations

Florian Lesaint[1,2]*, Olivier Sigaud[1,2], Shelly B. Flagel[3,4,5], Terry E. Robinson[5], Mehdi Khamassi[1,2]

1 Institut des Systèmes Intelligents et de Robotique, UMR 7222, UPMC Univ Paris 06, Paris, France, **2** Institut des Systèmes Intelligents et de Robotique, UMR 7222, CNRS, Paris, France, **3** Department of Psychiatry, University of Michigan, Ann Arbor, Michigan, United States of America, **4** Molecular and Behavioral Neuroscience Institute, University of Michigan, Ann Arbor, Michigan, United States of America, **5** Department of Psychology, University of Michigan, Ann Arbor, Michigan, United States of America

Abstract

Reinforcement Learning has greatly influenced models of conditioning, providing powerful explanations of acquired behaviour and underlying physiological observations. However, in recent autoshaping experiments in rats, variation in the form of Pavlovian conditioned responses (CRs) and associated dopamine activity, have questioned the classical hypothesis that phasic dopamine activity corresponds to a reward prediction error-like signal arising from a classical Model-Free system, necessary for Pavlovian conditioning. Over the course of Pavlovian conditioning using food as the unconditioned stimulus (US), some rats (sign-trackers) come to approach and engage the conditioned stimulus (CS) itself – a lever – more and more avidly, whereas other rats (goal-trackers) learn to approach the location of food delivery upon CS presentation. Importantly, although both sign-trackers and goal-trackers learn the CS-US association equally well, only in sign-trackers does phasic dopamine activity show classical reward prediction error-like bursts. Furthermore, neither the acquisition nor the expression of a goal-tracking CR is dopamine-dependent. Here we present a computational model that can account for such individual variations. We show that a combination of a Model-Based system and a revised Model-Free system can account for the development of distinct CRs in rats. Moreover, we show that revising a classical Model-Free system to individually process stimuli by using factored representations can explain why classical dopaminergic patterns may be observed for some rats and not for others depending on the CR they develop. In addition, the model can account for other behavioural and pharmacological results obtained using the same, or similar, autoshaping procedures. Finally, the model makes it possible to draw a set of experimental predictions that may be verified in a modified experimental protocol. We suggest that further investigation of factored representations in computational neuroscience studies may be useful.

Citation: Lesaint F, Sigaud O, Flagel SB, Robinson TE, Khamassi M (2014) Modelling Individual Differences in the Form of Pavlovian Conditioned Approach Responses: A Dual Learning Systems Approach with Factored Representations. PLoS Comput Biol 10(2): e1003466. doi:10.1371/journal.pcbi.1003466

Editor: Olaf Sporns, Indiana University, United States of America

Received July 1, 2013; **Accepted** December 19, 2013; **Published** February 13, 2014

Funding: This work was supported by Grant ANR-11-BSV4-006 "LU2" (Learning Under Uncertainty) from L'Agence Nationale de la Recherche, France (FL, OS, MK), by Grant "HABOT" from the Ville de Paris Emergence(s) Program, France (MK), by Grant "GoHaL" from the Centre National de la Recherche Scientifique PEPS Program, France (MK), and by Grant P01 DA031656 from the National Institute on Drug Abuse, USA (SBF, TER). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: lesaint@isir.upmc.fr

Introduction

Standard Reinforcement Learning (RL) [1] is a widely used normative framework for modelling conditioning experiments [2,3]. Different RL systems, mainly Model-Based and Model-Free systems, have often been combined to better account for a variety of observations suggesting that multiple valuation processes coexist in the brain [4–6]. Model-Based systems employ an explicit model of consequences of actions, making it possible to evaluate situations by forward inference. Such systems best explain goal-directed behaviours and rapid adaptation to novel or changing environments [7–9]. In contrast, Model-Free systems do not rely on internal models and directly associate values to actions or states by experience such that higher valued situations are favoured. Such systems best explain habits and persistent behaviours [9–11]. Of significant interest, learning in Model-Free systems relies on a computed reinforcement signal, the reward prediction error (RPE). This signal parallels the observed shift of dopamine neurons' response from the time of an initially unexpected reward – an outcome that is better or worse than expected – to the time of the conditioned stimulus that precedes it, which, in Pavlovian conditioning experiments, is fully predictive of the reward [12,13].

However recent work by Flagel et al. [14], raises questions about the exclusive use of classical RL Model-Free methods to account for data in Pavlovian conditioning experiments. Using an autoshaping procedure, a lever-CS was presented for 8 seconds, followed immediately by delivery of a food pellet into an adjacent food magazine. With training, some rats (sign-trackers; STs) learned to rapidly approach and engage the lever-CS. However, others (goal-trackers; GTs) learned to approach the food magazine

Author Summary

Acquisition of responses towards full predictors of rewards, namely Pavlovian conditioning, has long been explained using the reinforcement learning theory. This theory formalizes learning processes that, by attributing values to situations and actions, makes it possible to direct behaviours towards rewarding objectives. Interestingly, the implied mechanisms rely on a reinforcement signal that parallels the activity of dopamine neurons in such experiments. However, recent studies challenged the classical view of explaining Pavlovian conditioning with a single process. When presented with a lever whose retraction preceded the delivery of food, some rats started to chew and bite the food magazine whereas others chew and bite the lever, even if no interactions were necessary to get the food. These differences were also visible in brain activity and when tested with drugs, suggesting the coexistence of multiple systems. We present a computational model that extends the classical theory to account for these data. Interestingly, we can draw predictions from this model that may be experimentally verified. Inspired by mechanisms used to model instrumental behaviours, where actions are required to get rewards, and advanced Pavlovian behaviours (such as overexpectation, negative patterning), it offers an entry point to start modelling the strong interactions observed between them.

upon CS presentation, and made anticipatory head entries into it. Furthermore, in STs, phasic dopamine release in the nucleus accumbens, measured with fast scan cyclic voltammetry, matched RPE signalling, and dopamine was necessary for the acquisition of a sign-tracking CR. In contrast, despite the fact that GTs acquired a Pavlovian conditioned approach response, this was not accompanied with the expected RPE-like dopamine signal, nor was the acquisition of a goal-tracking CR blocked by administration of a dopamine antagonist (see also [15]).

Classical dual systems models [16–19] should be able to account for these behavioural and pharmacological data, but the physiological data are not consistent with the classical view of RPE-like dopamine bursts. Based on the observation that STs and GTs focus on different stimuli in the environment, we suggest that the differences observed in dopamine recordings may be due to an independent valuation of each stimulus. In classical RL, valuation is usually done at the *state* level. Stimuli, embedded into *states* – snapshots of specific configurations in time –, are therefore hidden to systems. In this case, it would prevent dealing separately with the lever and the magazine at the same time. However, such data may still be explained by a dual systems theory, when extended to support and benefit from factored representations; that is, learning the specific value of stimuli independently from the states in which they are presented.

In this paper, we present and test a model using a large set of behavioural, physiological and pharmacological data obtained from studies on individual variation in Pavlovian conditioned approach behaviour [14,20–25]. It combines Model-Free and Model-Based systems that provide the specific components of the observed behaviours [26]. It explains why inactivating dopamine in the core of the nucleus accumbens or in the entire brain results in blocking specific components and not others [14,25]. By weighting the contribution of each system, it also accounts for the full spectrum of observed behaviours ranging from one extreme – sign-tracking – to the other [26] – goal-tracking. Above all, by extending classical Model-Free methods with factored representations, it potentially explains why the lever-CS and the food magazine might acquire different motivational values in different individuals, even when they are trained in the same task [22]. It may also account for why the RPE-like dopaminergic responses are observed in STs but not GTs, and also the differential dependence on dopamine [14].

Results

We model the task as a simple Markov Decision Process (MDP) with different paths that parallel the diverse observed behaviours ranging from sign-tracking – engaging with the lever as soon as it appears – to goal-tracking – engaging with the magazine as soon as the lever-CS appears – (see Figure 1).

The computational model (see Figure 2) consists of two learning systems, employing distinct mechanisms to learn the same task: (1)

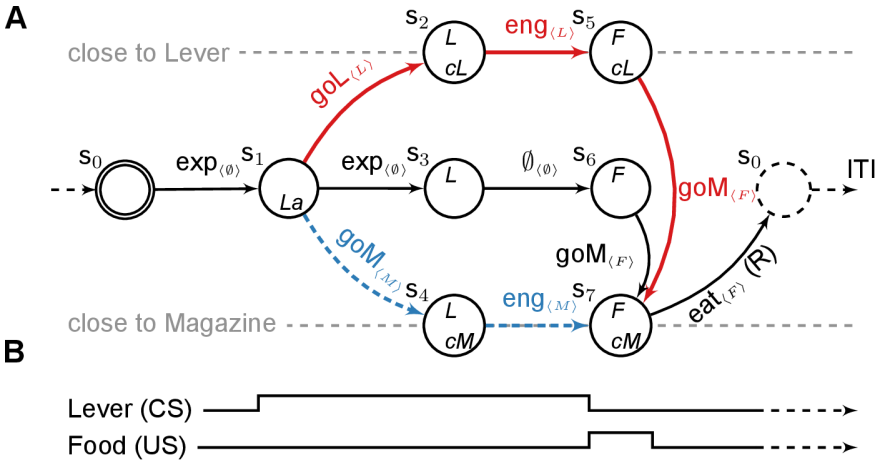


Figure 1. Computational representation of the autoshaping procedure. (**A**) MDP accounting for the experiments described in [14,21,22,26]. States are described by a set of variables: *L*/*F* - Lever/Food is available, *cM*/*cL* - close to the Magazine/Lever, *La* - Lever appearance. The initial state is double circled, the dashed state is terminal and ends the current episode. Actions are *eng*age with the proximal stimuli, *exp*lore, or *go* to the *M*agazine/*L*ever and *eat*. For each action, the feature that is being focused on is displayed within brackets. The path that STs should favour is in red. The path that GTs should favour is in dashed blue. (**B**) Time line corresponding to the unfolding of the MDP.
doi:10.1371/journal.pcbi.1003466.g001

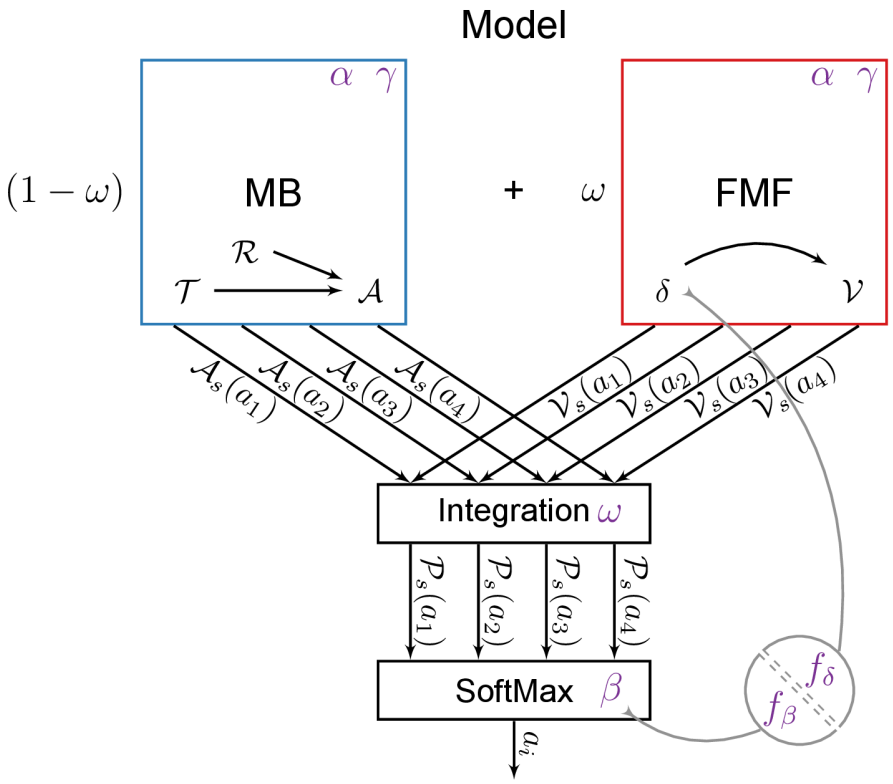


Figure 2. General architecture of the model and variants. The model is composed of a Model-Based system (MB, in blue) and a Feature-Model-Free system (FMF, in red) which provide respectively an Advantage function $\mathcal{A}$ and a value function $\mathcal{V}$ values for actions a_i given a state s. These values are integrated in $\mathcal{P}$, prior to be used into an action selection mechanism. The various elements may rely on parameters (in purple). The impact of flupentixol on dopamine is represented by a parameter f that influences the action selection mechanism and/or any reward prediction error that might be computed in the model.
doi:10.1371/journal.pcbi.1003466.g002

a Model-Based system which learns the structure of the task from which it infers its values; (2) a Feature-Model-Free system where values for the relevant stimuli (lever-CS and the food magazine) are directly learned by trial and error using RPEs. The respective values of each system are then weighted by an ω parameter before being used in a classical softmax action-selection mechanism (see Methods).

An important feature of the model is that varying the systems weighting parameter ω (while sharing the other parameter values of the model across subgroups) is sufficient to qualitatively reproduce the characteristics of the different subgroups of rats observed experimentally during these studies.

To improve the matching of the following results with the main experimental data, a different set of parameter values was used for each subgroup (ST, GT and IG). The values were retrieved after fitting autoshaping data only (see Methods, Table S1). Simulated results on other behavioural, physiological and pharmacological data are generated with the same parameter values. While it might result in a weaker fitting of the other experimental data, this permits a straightforward comparison of results at different levels for the same simulation. Moreover, it confirms that the model can reproduce behavioural, physiological and pharmacological results with a single simulation per subgroup.

On each set of experimental data, we compare different variants of the computational model in order to highlight the key mechanisms that are required for their reproduction. Simulation results on each data subset are summarized in Figure 3. The role of each specific mechanism of the model in reproducing each experimental data is detailed in Figure 4.

Behavioural data

Autoshaping. The central phenomenon that the model is meant to account for is the existence of individual behavioural differences in the acquisition of conditioned approach responses in rats undergoing an autoshaping procedure; that is, the development of a sign-tracking CR, a goal-tracking CR, or an intermediate response.

Based on their engagement towards the lever, Flagel et al. [21] divided rats into three groups (see [26] for a more recently defined criterion). At lever appearance, rats that significantly increased their engagement towards it (top 30%) were classified as STs, whereas rats that almost never engaged with the lever (bottom 30%) were classified as GTs (these latter animals engaged the food magazine upon CS presentation). The remaining rats, engaging in both lever and magazine approach behaviours were defined as the Intermediate Group (IGs) (see Figure 5 A, B). STs and GTs acquired their respective CRs at a similar rate over days of training [22].

The current model is able to reproduce such results (see Figure 5 C, D). By running a simulation for each group of rats, using different parameters (mainly varying the ω parameter) the model reproduces the different tendencies to engage with the lever ($\omega = 0.499$), with the magazine ($\omega = 0.048$) or to fluctuate between the two ($\omega = 0.276$). A high ω strengthens the influence of the Feature-Model-Free system, which learns to associate a high motivational value to the lever CS, and a sign-tracking CR dominates. A low ω increases the influence of the Model-Based system, which infers the optimal behaviour to maximize reward,

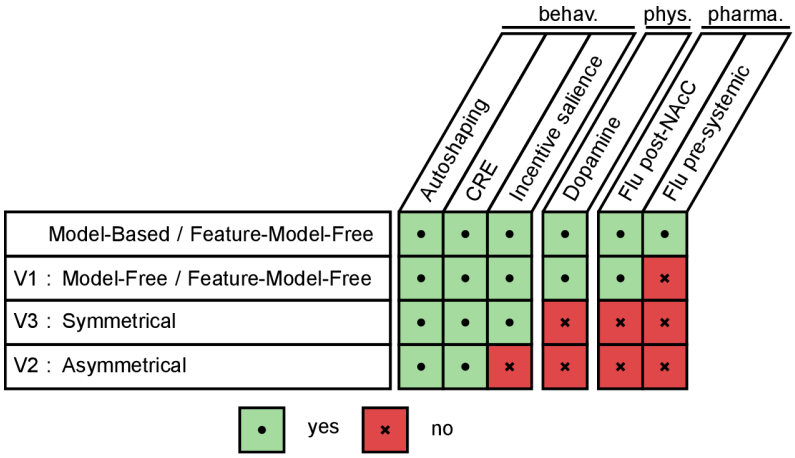


Figure 3. Summary of simulations and results. Each line represents a different model composed of a pair of Reinforcement Learning systems. Each column represents a simulated experiment. Experiments are grouped by the kind of data accounted for: behavioural (autoshaping [14,21], CRE [22], Incentive salience [23,24]), physiological [21] and pharmacological (Flu post-NAcC [25], Flu pre-systemic [21]). Variant 4 (i.e. Model-based/Model-Free without features) is not included as it failed to even reproduce the autoshaping behavioural results and was not investigated further.
doi:10.1371/journal.pcbi.1003466.g003

and goal-tracking is favoured. When both systems are mixed, i.e. with an intermediate ω, the behaviour is more likely to oscillate between sign- and goal-tracking, representative of the intermediate group.

These results rely on the combination of two systems that would independently lead to 'pure' sign-tracking or goal-tracking CRs. Three tested variants of the model could reproduce these behavioural results as well (see Figure S1): a combination of Feature-Model-Free systems and simple Model-Free system (Variant 1); a multi-step extension of Dayan 2006's model [16] giving a Pavlovian impetus for the lever (Variant 2); and a symmetrical version of this last model with two impetuses, one for the lever, and one for the magazine (Variant 3) (see Methods). Interestingly, a combination of Model-Based and classical Model-Free (not feature-based : Variant 4) fails in reproducing these results (see Figure S8). This is because both systems are proven to converge to the same values and both would favour pure goal-tracking, such that varying their contribution has no impact on the produced behaviours.

Thus, at this stage, we can conclude that several computational models based on dual learning systems can reproduce these behavioural results, given that the systems favour different

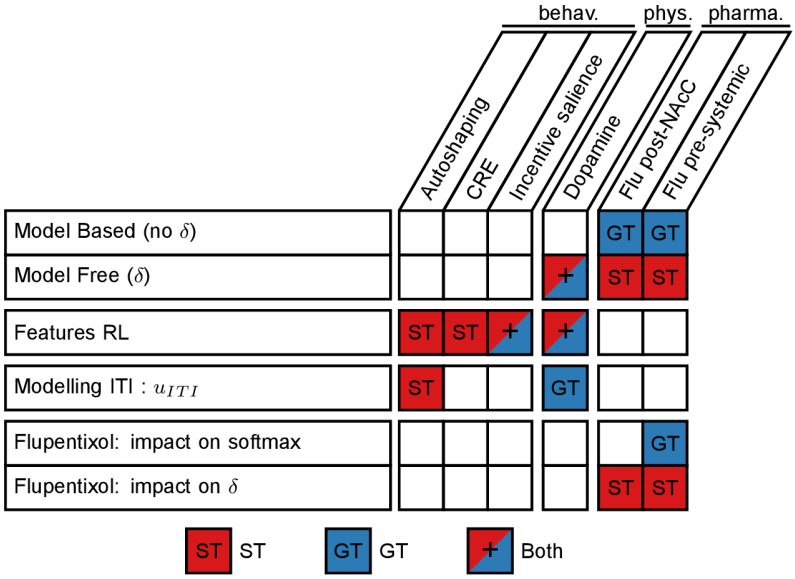


Figure 4. Summary of the key mechanisms required by the model to reproduce experimental results. Each line represents a different mechanism of the model. Each column represents a simulated experiment. For each mechanism, it states in which experiment and for which behaviour – sign-tracking (red), goal-tracking (blue) or both (+) – it is required. Note however that all mechanisms and associated parameters have, to a certain extent, an impact on any presented results.
doi:10.1371/journal.pcbi.1003466.g004

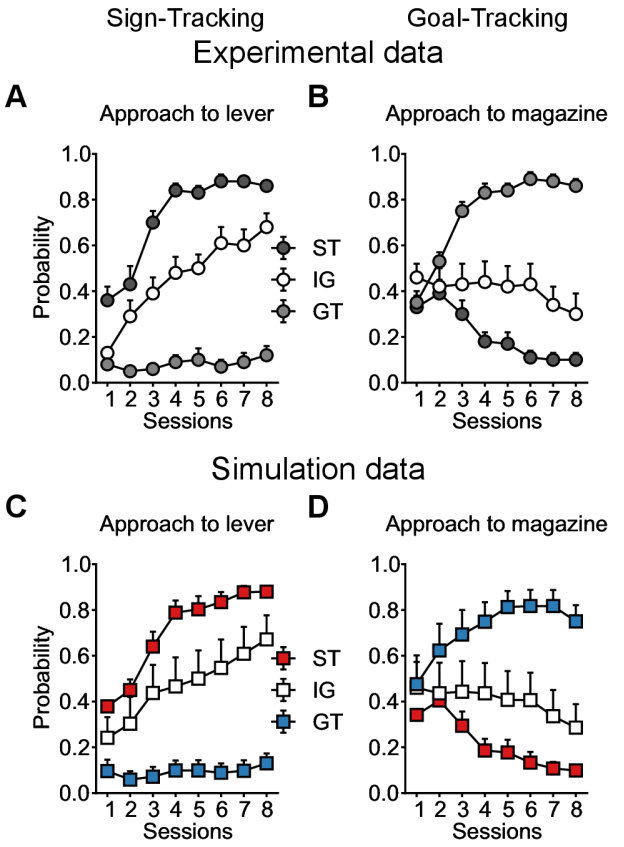


Figure 5. Reproduction of sign- versus goal-tracking tendencies in a population of rats undergoing an autoshaping experiment. Mean probabilities to engage at least once with the lever (**A,C**) or the magazine (**B,D**) during trials. Data are expressed as mean $\pm$ S.E.M. and illustrated in 50-trial (2-session) blocks. (**A,B**) Reproduction of Flagel et al. [21] experimental results (Figure 2 A,B). Sign-trackers (ST) made the most lever presses (black), goal-trackers (GT) made the least lever presses (white), Intermediate group (IG) is in between (grey). (**C,D**) Simulation of the same procedure (squares) with the model. Simulated groups of rats are defined as STs ($\omega=0.499$; $\beta=0.239$; $\alpha=0.031$; $\gamma=0.996$; $u_{ITI}=0.027$; $\mathcal{Q}_i(s_1,goL)=0.844$; $\mathcal{Q}_i(s_1,exp)=0.999$; $\mathcal{Q}_i(s_1,goM)=0.538$; n = 14) in red, GTs ($\omega=0.048$; $\beta=0.084$; $\alpha=0.895$; $\gamma=0.727$; $u_{ITI}=0.140$; $\mathcal{Q}_i(s_1,goL)=1.0$; $\mathcal{Q}_i(s_1,exp)=0.316$; $\mathcal{Q}_i(s_1,goM)=0.023$; n = 14) in blue and IGs ($\omega=0.276$; $\beta=0.142$; $\alpha=0.217$; $\gamma=0.999$; $u_{ITI}=0.228$; $\mathcal{Q}_i(s_1,goL)=0.526$; $\mathcal{Q}_i(s_1,exp)=0.888$; $\mathcal{Q}_i(s_1,goM)=0.587$; n = 14) in white. The model reproduces the same behavioural tendencies. With training, STs tend to engage more and more with the lever and less with the magazine, while GTs neglect the lever to increasingly engage with the magazine. IGs are in between.
doi:10.1371/journal.pcbi.1003466.g005

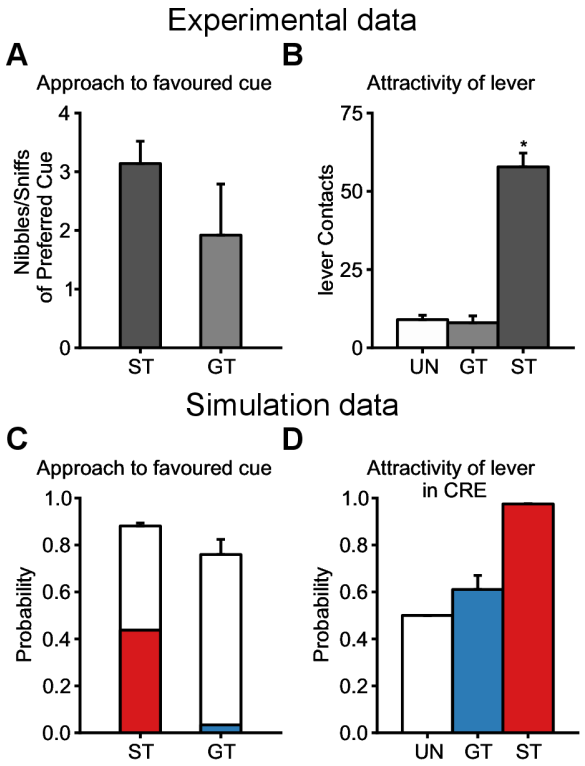


Figure 6. Possible explanation of incentive salience and Conditioned Reinforcement Effect by values learned during autoshaping procedure. Data are expressed as mean $\pm$ S.E.M. Simulated groups of rats are defined as in Figure 5. (**A**) Number of nibbles and sniffs of preferred cue by STs and GTs as a measure for incentive salience. Data extracted from Mahler et al. [23] from Figure 3 (bottom-left). (**B**) Reproduction of Robinson et al. [22] experimental results (Figure 2 B). Lever contacts by STs and GTs during a conditioned reinforcer experiment. (**C**) Probability to engage with the respective favoured stimuli of STs and GTs at the end of the simulation (white, similar to the last session of Figure 5 C for STs and D for GTs) superimposed with the contribution in percentage of the values attributed by the Feature-Model-Free system in such engagement for STs (red) and GTs (blue). We hypothesize that such value is the source of incentive salience and explains why STs and GTs have a consumption-like behaviour towards their favoured stimulus. (**D**) Probability to engage with the lever versus exploring when presented with the lever and no magazine for STs (red), GTs (blue) and a random-policy group UN (white), simulating the unpaired group (UN) of the experimental data. Probabilities were computed by applying the softmax function after removing the values for the magazine interactions (see Methods). STs would hence actively seek to engage with the lever relatively to GTs in a Conditioned Reinforcement Effect procedure.
doi:10.1371/journal.pcbi.1003466.g006

behaviours (see Figure S1). However, Variants 1, 2 and 3 fail to reproduce other behavioural, pharmacological and physiological data characteristic of STs and GTs (see following sections).

Incentive salience. The results in Figure 5 only represent the probability of approach to either the lever-CS or the food magazine. Thus, they do not account for the specific ways rats engage and interact with the respective stimuli. In fact, if food is used as the US, rats are known to chew and bite the stimuli on which they are focusing [23,24] (see Figure 6 A). Importantly, both STs and GTs express this consumption-like behaviour during the CS period, directed towards the lever or the food magazine, respectively. It has been argued that this behaviour may reflect the degree to which incentive salience is attributed to these stimuli, and thus the extent to which they become "wanted" [23,24,27].

In an RL-like framework, incentive salience attribution can be represented as a bonus mechanism for interacting with stimuli. The Feature-Model-Free system in the model realizes such a function, providing a specific bonus for each stimulus in any simulated rat. Such bonus was inspired by the Pavlovian impetus mechanism of Dayan 2006's model [16]. Figure 6 C shows the percentage of Feature-Model-Free value that contributed to the computation of the probability to engage with the respective favoured cues of STs and GTs at the end of the simulation.

The presence of the magazine in the inter-trial interval (ITI), and the necessary revision of the associated bonus at a lower value

when exploring, makes the associated bonus smaller than that of the lever (see Methods). This results in a even smaller contribution of this bonus in GTs behaviour (blue bar in Figure 6 C) compared to STs (red bar in Figure 6 C). Although it is not straightforward to interpret how the probability of engagement (white bars in Figure 6 C) in the model might be translated into a consumption-like behaviour from a computational point of view, we propose that the different contributions of bonuses could explain the slightly smaller number of nibbles and sniffs of preferred cue observed experimentally in GTs compared to STs (Figure 6 A, adapted from [23]). This may also explain why other studies have observed a smaller proportion of nibbles on the magazine in GTs [24] and less impulsiveness [28] in GTs compared to STs. We come back to this issue in the discussion.

Variants 1 and 3 also realize such function by providing bonuses for actions leading to both stimuli (see Figure S2). Only providing bonus for sign-tracking behaviour – as in Dayan's model (Variant 2) – does not fit well with the attribution of incentive salience to both stimuli. It would suggest that we should not observe incentive salience towards the magazine in any rats, which is in discrepancy with the experimental data. Thus, the important mechanism here is that stimuli are not processed differently. Any stimulus is attributed with its respective bonus, which is pertinent in regard to the attribution of incentive salience.

Conditioned Reinforcement Effect (CRE). An important question about the difference in observed behaviours is about the properties acquired by the lever that makes it more attractive to STs than to GTs. To answer this question, Robinson and Flagel studied the dissociation of the predictive and motivational properties of the lever [22]. Part of their results involves asking whether the Pavlovian lever-CS would serve as a conditioned reinforcer, capable of reinforcing the learning of a new instrumental response [29,30]. In a new context, rats were presented with an active and an inactive nose port. Nose poking into the active port resulted in presentation of the lever for 2 seconds without subsequent reward delivery, whereas poking into the inactive one had no consequence. The authors observed that while both STs and GTs preferred the active nose port to an inactive one, STs made significantly more active nose pokes than GTs (see Figure 6 B, see also [31]). This suggests that the lever acquired greater motivational value in STs than in GTs.

Without requiring additional simulations, the model can explain these results by the value that has been incrementally learned and associated with approaching the lever in the prior autoshaping procedure for STs and GTs. In the model, STs attribute a higher value to interacting with the lever than GTs and should actively work for its appearance enabling further engagement. Figure 6 D shows the probabilities of engagement that would be computed at lever appearance after removing the magazine (and related actions) at the end of the experiment. Indeed, even though the lever is presented only very briefly, upon its presentation in the conditioned reinforcement test, STs actively engage and interact with it [22]. Any value associated to a state-action pair makes this action in the given state rewarding in itself, favouring actions (e.g. nosepokes) that would lead to such state. Repeatedly taking this action without receiving rewards should eventually lead to a decrease of this value and reduce the original engagement.

Physiological data

Not only have Flagel et al. [14] provided behavioural data but they also provide physiological and pharmacological data. This raises the opportunity to challenge the model at different levels, as developed in the current and next sections.

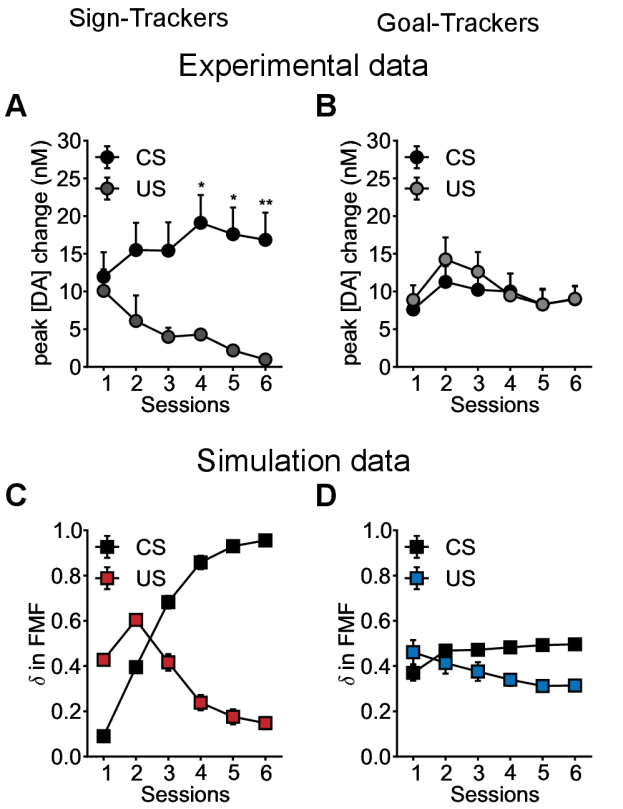


Figure 7. Reproduction of patterns of dopaminergic activity of sign- versus goal-trackers undergoing an autoshaping experiment. Data are expressed as mean ± S.E.M. (**A,B**) Reproduction of Flagel et al. [14] experimental results (Figure 3 d,f). Phasic dopamine release recorded in the core of the nucleus accumbens in STs (light grey) and GTs (grey) using Fast Scan Cyclic Voltammetry. Change in peak amplitude of the dopamine signal observed in response to CS and US presentation for each session of conditioning (**C,D**) Average RPE computed by the Feature-Model-Free system in response to CS and US presentation for each session of conditioning. Simulated groups of rats are defined as in Figure 5. The model is able to qualitatively reproduce the physiological data. STs (blue) show a shift of activity from US to CS time over training, while GTs develop a second activity at CS time while maintaining the initial activity at US time.
doi:10.1371/journal.pcbi.1003466.g007

Using Fast Scan Cyclic Voltammetry (FSCV) in the core of the nucleus accumbens they recorded the mean of phasic dopamine (DA) signals upon CS (lever) and US (food) presentation. It was observed that depending on the subgroup of rats, distinct dopamine release patterns emerge (see Figure 7 A,B) during Pavlovian training. STs display the classical propagation of a phasic dopamine burst from the US to the CS over days of training and the acquisition of conditioned responding (see Figure 7 A). This pattern of dopamine activity is similar to that seen in the firing of presumed dopamine cells in monkeys reported by Schultz and colleagues [12] and interpreted as an RPE corresponding to the reinforcement signal δ of Model-Free RL systems [1]. In GTs, however, a different pattern was observed. Initially there were small responses to both the CS and US, of which the amplitudes seemed to follow a similar trend over training (see Figure 7 B).

By recording the mean of the RPEs δ computed in the Feature-Model-Free system during the autoshaping simulation (i.e. only fitted to behavioural data), the model can still qualitatively

reproduce the different patterns observed in dopamine recordings for STs and GTs (see Figure 7 C,D). For STs, the model reproduces the progressive propagation of δ from the US to the CS (see Figure 7 C). For GTs, it reproduces the absence of such propagation. The RPE at the time of the US remains over training, while a δ also appears at the time of the CS (see Figure 7 D). In the model, such discrepancy is explained by the difference in the values that STs and GTs use for the computation of RPEs at the time of the CS and the US. STs, by repeatedly focusing on the lever, propagate the total value of food to the lever and end up having a unique δ at the unexpected lever appearance only. By contrast, by repeatedly focusing on the magazine during the lever appearance but, as all rats, also from time to time during ITI, GTs revise the magazine value multiple times, positively just after food delivery and negatively during ITI. Such revisions lead to a permanent discrepancy between the expected and observed value, i.e. a permanent δ, at lever appearance and food delivery, when engaging with the magazine.

The key mechanism to reproduce these results resides in the generalization capacities of the Feature-Model-Free system. Based on features rather than states, feature-values are to be used, and therefore revised, at different times and states of the experiment, favouring the appearance of RPEs. Variants 2, 3 and 4 relying on classical Model-Free systems are unable to reproduce such results (see Figure S3). By using values over abstract states rather than stimuli, it makes it impossible to only revise the value of the magazine during ITI. Therefore, given the deterministic nature of the MDP, we observe a classical propagation of RPEs in all pathways up to the appearance of the lever.

Pharmacological data

Effects of systemic flupentixol administration on the learning of sign- and goal-tracking behaviours. Flagel et al. [14] also studied the impact of systemic injections of the non specific dopamine antagonist, flupentixol, on the acquisition of sign-tracking and goal-tracking CRs. The authors injected flupentixol in rats prior to each of 7 sessions and observed the resulting behaviours. Behaviour during the 8^{th} session was observed without flupentixol.

Systemic injections of flupentixol in STs and GTs (Flu groups, black curves in Figure 8 A,B) blocked expression of their respective behaviours during training. Saline injections (white curves in Figure 8 A,B) left their performances intact. The crucial test for learning took place on the 8^{th} day, when all rats were tested without flupentixol. STs failed to approach the lever, and performed as the saline-injected controls did on the first day of training.

Thus, in STs flupentixol blocked the acquisition of a sign-tracking CR (see Figure 8 A). Interestingly, on the flupentixol-free test day GTs did not differ from the saline-injected control group, indicating that flupentixol did not block the acquisition of a goal-tracking CR (see Figure 8 B). Thus, acquisition of a sign-tracking CR, but not a goal-tracking CR, is dependent on dopamine (see also [15]).

The model reproduces these pharmacological results (see Figure 8 C,D). As in the experimental data, simulated GTs and STs do not show a specific conditioned response during the first 7 sessions under flupentixol. On the 8^{th} session, without flupentixol, we observe that STs still do not show a specific conditioned response while GTs perform at a level close to that of the saline-injected control group (see Figure 8 C,D).

The absence of specific conditioned response in the whole population for the first 7 sessions is first due to the hypothesized [32] impact of flupentixol on action selection (see Methods). With enough flupentixol, the elevation of the selection temperature leads to a decrease of the influence of learned values in the expressed behaviour, masking any possibly acquired behaviour.

The absence of a specific conditioned response in STs is due to the blockade of learning in the second system by flupentixol, since it is RPE-dependent. Therefore almost no learning occurs in the system (see Figure 8).

In contrast, with the first system being RPE-independent, flupentixol has no effect on learning, because it is Model-Based rather than Model-Free [33]. The expression of behaviour is blocked at the action selection level, which does not make use of values learned by the Model-Based system. Thus, GTs, relying mainly on the first system, learn their CR under flupentixol but are just not able to express it until flupentixol is removed. The lower level of goal-tracking in the Flu group relative to the saline-injected control group on the 8^{th} session is due to the lack of exploitation induced by flupentixol injection during the previous 7 sessions. By engaging less with the magazine, the Flu group ends up associating a lower value to the magazine (i.e. the value did not fully converge in 7 sessions) to guide its behaviour.

Interestingly, if the model had been constituted of Model-Free systems only – as in Variants 1, 2 and 3 – it would not have been able to reproduce these results, because both systems would have been RPE-dependent and thus sensitive to the effect of flupentixol (see Figure S4).

Effects of local flupentixol administration on the expression of sign- and goal-tracking behaviours. In a related experiment, Saunders et al. [25] studied the role of dopamine in the nucleus accumbens core in the expression of Pavlovian-conditioned responses that had already been acquired. After the same autoshaping procedure as in [20], they injected different doses of flupentixol in the core of the nucleus accumbens of rats and quantified its impact on the expression of sign-tracking and goal-tracking CRs in an overall population (without distinguishing between STs and GTs).

They found that flupentixol dose dependently attenuated the expression of sign-tracking, while having essentially no effect on goal-tracking (see Figure 9 A, B). Along with the Flagel et al. [14] study, these results suggest that both the acquisition and expression of a sign-tracking CR is dopamine-dependent (at least in the core) whereas the acquisition and expression of a goal-tracking CR is not.

Given the assumption that the Feature-Model-Free system would take place in or rely on the core of the nucleus accumbens, this model reproduces the main experimental result: the decreased tendency to sign-track in the population (see Figure 9 C). Note that in the previous experiment, the injection of flupentixol was systemic, and assumed to affect any region of the brain relying on dopamine, whereas in the present experiment it was local to the core of the nucleus accumbens. Therefore, we modelled the impact of flupentixol differently between the current and previous simulations (see Methods). In the model, the tendency to sign-track is directly correlated with a second operational system. Any dysfunction in the learning process (here by a distortion of RPEs) reduces this trend.

The model successfully reproduced the absence of reduction of goal-tracking, in contrast to the reduction of sign-tracking. However, it was unable to reproduce the invariance in goal-tracking (see Figure 9 D) and rather produced an increase in goal-tracking. This is due to the use of a softmax operator for action selection, as this is the case in the vast majority of computational neuroscience RL models [16–19,32,34–36], which automatically favours goal-tracking when sign-tracking is blocked (see Limitations). We did not attempt to cope with this limitation because our focus here was the absence of reduction of goal-tracking.

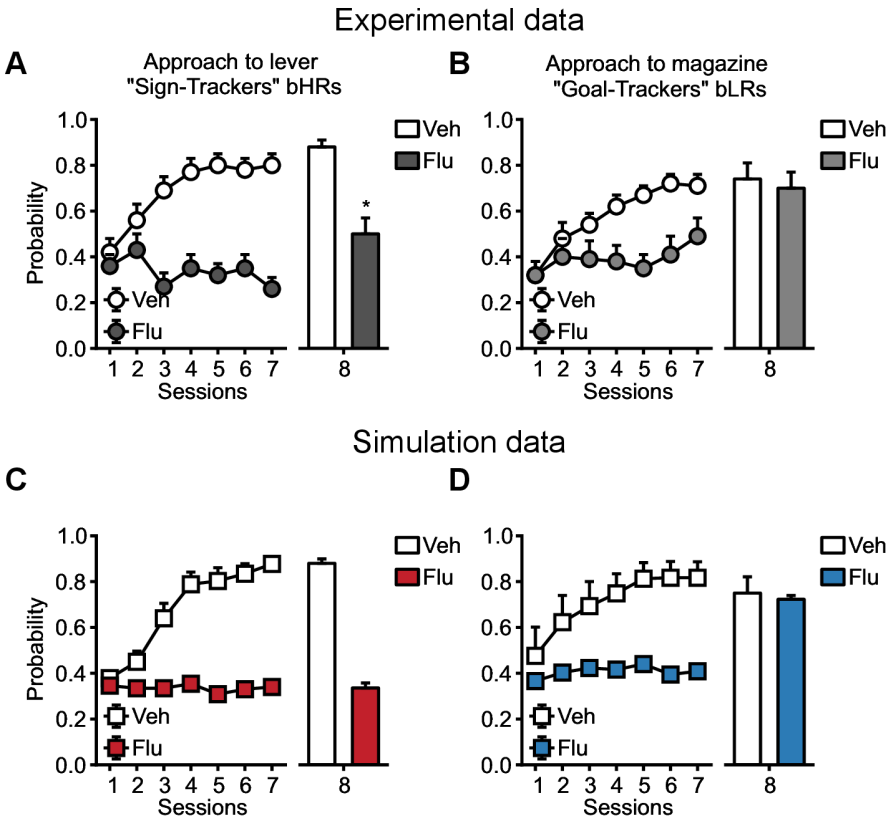


Figure 8. Reproduction of the effect of systemic injections of flupentixol on sign-tracking and goal-tracking behaviours. Data are expressed as mean ± S.E.M. (**A,B**) Reproduction of Flagel et al. [14] experimental results (Figure 4 a,d). Effects of flupentixol on the probability to approach the lever for STs (**A**) and the magazine for GTs (**B**) during lever presentation. (**C,D**) Simulation of the same procedure (squares) with the model. Simulated groups of rats are defined as in Figure 5. (**C**) By flattening the softmax temperature and reducing the RPEs of the Feature-Model-Free system, to mimic the possible effect of flupentixol, the model can reproduce the blocked acquisition of sign-tracking in STs (red), engaging less the lever relatively to a saline-injected control group (white). (**D**) Similarly, the model reproduces that goal-tracking was learned but its expression was blocked. Under flupentixol (first 7 sessions), GTs (blue) did not express goal-tracking, but on a flupentixol-free control test (8^{th} session) their engagement with the magazine was almost identical to the engagement of a saline-injected control group (white).
doi:10.1371/journal.pcbi.1003466.g008

Besides, the model could, after re-learning, reproduce the selective impact of intra-accumbal flupentixol injections observed in sign-tracking but not in goal-tracking, because such injections affected the learning process in the Feature-Model-Free system only.

Discussion

We tested several mechanisms from the current literature on modelling individual variation in the form of Pavlovian conditioned responses (ST vs GT) that emerge using a classical autoshaping procedure, and the role of dopamine in both the acquisition and expression of these CRs. Benefiting from a rich set of data, we identified key mechanisms that are sufficient to account for specific properties of the observed behaviours. The resulting model relies on two major concepts: Dual learning systems and factored representations. Figure 4 summarizes the role of each mechanism in the model.

Dual learning systems

Combining Model-Based and Model-Free systems has previously been successful in explaining the shift from goal-directed to habitual behaviours observed in instrumental conditioning [17–19,33,34]. However, few models based on the same concept have been developed to account for Pavlovian conditioning [16]. While the need for two systems is relevant in instrumental conditioning given the distinct temporal engagement of each system, such a distinction has not been applied to Pavlovian phenomena (but see recent studies on orbitofrontal cortex [37–39]). The variability of behaviours and the need for multiple systems have been masked by focusing on whole populations and, for the most part, ignoring individual differences in studies of Pavlovian conditioning. The nature of the CS is especially important, as many studies of Pavlovian conditioned approach behaviour have used an auditory stimulus as the CS, and in such cases only a goal-tracking CR emerges in rats [40,41].

As expected from the behavioural data, combining two learning systems was successful in reproducing sign- and goal-tracking behaviours. The Model-Based system, learning the structure of the task, favours systematic approach towards the food magazine, and waiting for food to be delivered, and hence the development of a goal-tracking CR. The Feature-Model-Free system, directly evaluating features by trials and errors, favours systematic approach towards the lever, a full predictor of food delivery, and hence the development of a sign-tracking CR. Moreover, utilizing the Feature-Model-Free system to represent sign-tracking behaviour yields results consistent with the pharmacological data. Disrupting RPEs, which reflects the effects of flupentixol on

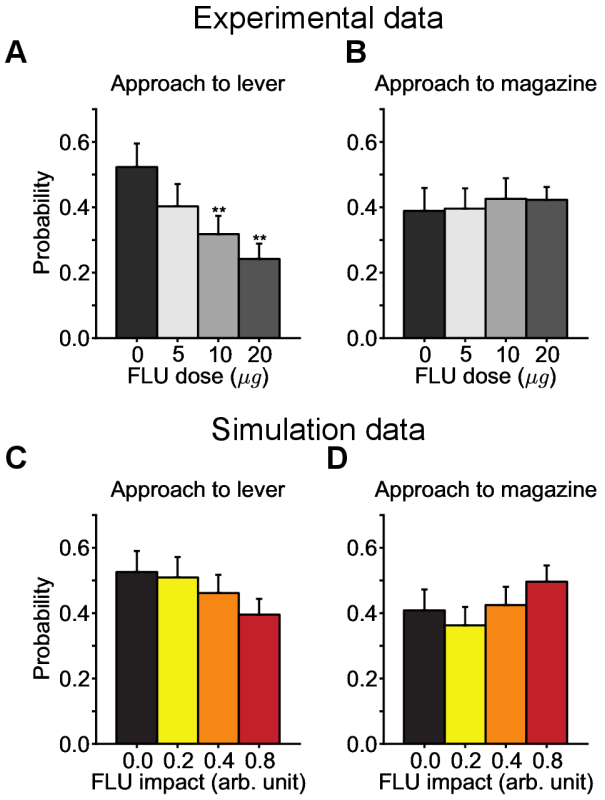


Figure 9. Reproduction of the effect of post injections of flupentixol in the core of the nucleus accumbens. Data are expressed as mean ± S.E.M. (**A,B**) Reproduction of Saunders et al. [25] experimental results (Figure 2 A,D). Effects of different doses of flupentixol on the general tendency to sign-track (**A**) and goal-track (**B**) in a population of rats, without discriminating between sign- and goal-trackers. (**C,D**) Simulation of the same procedure with the model. The simulated population is composed of groups of rats defined as in Figure 5. By simulating the effect of flupentixol as in Figure 8, the model is able to reproduce the decreasing tendency to sign-track in the overall population by increasing the dose of flupentixol.
doi:10.1371/journal.pcbi.1003466.g009

dopamine, blocks the acquisition of a sign-tracking CR, but not a goal-tracking CR. The model does not make a distinction between simple approach behaviour versus consumption-like engagement, as reported for both STs and GTs [23,24]. However given that such engagement results from the development of incentive salience [23,24], the values learned by the Feature-Model-Free system to bias behaviour towards stimuli attributed with motivational value are well-suited to explain such observations. The higher motivational value attributed to the lever by STs relative to GTs can also explain why the lever-CS is a more effective conditioned reinforcer for STs than for GTs [22].

Importantly, none of the systems are dedicated to a specific behaviour, nor rely on *a priori* information to guide their processes. The underlying mechanisms increasingly make one behaviour more pronounced than the other through learning. Each system contributes to a certain extent to sign- and goal-tracking behaviour. This property is emphasized by the weighted sum integration of the values computed by each system before applying the softmax action-selection mechanism. The variability of behaviours in the population can then be accounted for by adjusting the weighting parameter ω from 1 (i.e. favouring sign-tracking) to 0 (i.e. favouring goal-tracking). This suggests that the rats' actions result from some combination of rational and impulsive processes, with individual variation contributing to the weight of each component.

The integration mechanism is directly inspired by the work of Dayan et al. [16] and as the authors suggest, the parameter ω may fluctuate over time, making the contribution of the two systems vary with experience. In contrast to their model, however, the model presented here does not assign different goals to each system. Thus, the current model is more similar to their previous model [17], which uses another method for integration.

A common alternative to integration when using multiple systems [17,18,35] is to select at each step, based on a given criterion (certainty, speed/accuracy trade-off, energy cost), a single system to pick the next action. Such switch mechanism does not fit well with the present model, given that it would be interpreted as if actions relied sometimes only on motivational values (i.e. Feature-Model-Free system) and sometimes only on a rational analysis of the situation (i.e. Model-Based system). It also does not fit well with pharmacological observation that STs do not express goal-tracking tendencies in the drug-free test session following systemic-injections of flupentixol [14], as Flagel et al. stated, "[sign-tracking] rats treated with flupentixol did not develop a goal-tracking CR".

Factored representations

Classical RL algorithms used in neuroscience [16–18,35], designed mainly to account for instrumental conditioning, work at the state level. Tasks are defined as graphs of states, and corresponding models are unaware of any similarity within states. Therefore, any subsequent valuation process cannot use any underlying structure to generalize updates to states that share stimuli. Revising the valuation process to handle features rather than states *per se*, makes it possible to attribute motivational values to stimuli independently of the states in which they are presented.

Recent models dedicated to Pavlovian conditioning [36,42–46] usually represent and process stimuli independently and can be said to use factored representations, a useful property to account for phenomena such as blocking [47] or overexpectation [48]. In contrast to the present model, while taking inspiration from RL theory (e.g. using incremental updates), these models are usually far from the classical RL framework. Of significant difference with the present study, most of these models tend to describe the varying intensity of a unique conditioned response and do not account for variations in the actual form of the response, as we do here. In such models, the magazine would not be taken into account and/or taken as part of the context, making it unable to acquire a value for itself nor be the focus of a particular response.

In RL theory, factorization is mainly evoked when trying to overcome the curse of dimensionality [49] (i.e. standard algorithms do not scale well to high dimensional spaces and require too much physical space or computation time). Amongst methods that intend to overcome this problem are value function approximations and Factored Reinforcement Learning. Value function approximations [35,50,51] attempt to split problems into orthogonal subproblems making computations easier and providing valuations that can then be aggregated to estimate the value of states. Factored Reinforcement Learning [52–54] attempts to find similarities between states so that they can share values, reducing the physical space needed and relies on factored Markov Decision Processes. We also use factored Markov Decision processes, hence the "factored" terminology. However, our use of factored representations serves a different purpose. We do not intend to build a compact value-function nor infer the value of states from

values of features but rather make these values compete in the choice for the next action.

Taking advantage of factored representations into classical RL algorithms is at the very heart of the present results. By individually processing stimuli within states (i.e. in the same context, at the same time and same location) and making them compete, the Feature-Model-Free system favours a different policy – oriented towards engaging with the most valued stimuli – (sign-tracking) than would have been favoured by classical algorithms such as Model-Based or Model-Free systems (goal-tracking). Hence, combining a classical RL algorithm with the Feature-Model-Free system enables the model to reproduce the difference in behaviours observed between STs and GTs during an autoshaping procedure. Moreover, by biasing expected optimal behaviours towards cues with motivational values (incentive salience), it is well suited to explain the observed commitment to unnecessary and possibly counter-productive actions (see also [16,55,56]). Most of all, it enables the model to replicate the different patterns of dopamine activity recorded with FSCV in the core of the nucleus accumbens of STs and GTs. The independent processing of stimuli leads to patterns of RPE that match those of dopamine activity for STs – a shift of bursts from the US to the CS; and in GTs – a persistence of bursts at both the time of the US and the CS.

A promising combination

By combining the two concepts of dual learning systems and factored representations in a single model, we are able to reproduce individual variation in behavioural, physiological and pharmacological effects in rats trained using an autoshaping procedure. Interestingly, our approach does not require a deep revision of mechanisms that are extensively used in our current field of research.

While Pavlovian and instrumental conditioning seem entangled in the brain [57], the two major concepts on which rely their respective models, dual learning systems and factored representations, have to our knowledge never been combined into a single model in this field of research.

This approach could contribute to the understanding of interactions between these two classes of learning, such as CRE or Pavlovian-Instrumental Transfer (PIT), where motivation for stimuli acquired via Pavlovian learning modulates the expression of instrumental responses. Interestingly, the Feature-Model-Free system nicely fits with what would be expected from a mechanism contributing to general PIT [58]. It is focused on values over stimuli without regard to their nature [58], it biases and interferes with some more instrumental processes [55,56,58] and it is hypothesized to be located in the core of the nucleus accumbens [58]. It would thus be interesting to study whether future simulations of the model could explain and help better formalize these aspects of PIT.

We do not necessarily imply that instrumental and Pavlovian conditioning might rely on a unique model. Rather, we propose that if they were the results of separated systems, they should somehow rely on similar representations and valuation mechanisms, given the strength of the observed interactions.

Theoretical and practical implications

The proposed model explains the persistent dopamine response to the US in GTs over days of training as a permanent RPE due to the revision of the magazine value during each ITI. Therefore, a prediction of the model is that shortening the ITI should reduce the amplitude of this burst (i.e. there should be less time to revise the value and reduce the size of the RPE); whereas increasing the ITI should increase the amplitude of this burst. Removing the food dispenser during ITI, similar to theoretically suppressing the ITI, should make this same burst disappear. Studying physiological data by grouping them given the duration of the preceding ITI might be sufficient, relatively to noise, to confirm that its duration impacts the amplitude of dopamine bursts. In the current experimental procedure, the ITI is indeed randomly picked in a list of values with an average of 90 sec. Moreover, reducing ITI duration should lead to an increase of the tendency to goal-track in the overall population. Indeed, with a higher value of the food magazine, the Feature-Model-Free system would be less likely to favour sign-tracking over goal-tracking CR. The resulting decrease in sign-tracking in the overall population would be consistent with findings of previous works [59–62], where a shorter ITI reduces the observed performance in the acquisition of sign-tracking CRs. Alternatively, it would also be interesting to examine the amplitude of dopamine bursts during the ITI (especially when exploring the food magazine), to determine whether or not physiological responses during this period affect the outcome of the conditioned response.

It would be interesting to split physiological data not only between STs and GTs but also between the stimuli on which the rats started and/or ended focusing on during CS presentation at each trial. This would help to confirm that the pattern of dopamine activity is indeed due to a separate valuation of each stimuli. We would predict that at the time of the US, dopamine bursts during engagement with the lever should be small relatively to dopamine bursts during engagement with the magazine. Moreover, comparing dopamine activity at the time of the CS when engaging with the lever versus the magazine could help elucidate which update mechanism is being used. If activity differs, this would suggest that the model should be revised to use SARSA-like updates, i.e. taking into account the next action in RPE computation. Such a question has already been the focus of some studies on dopamine activity [63–65].

There is no available experimental data for the phasic dopaminergic activity of the intermediate group. The model predicts that such a group would have a permanent phasic dopamine burst, i.e. RPE, at US and a progressively appearing burst at CS (see Figure S6). Over training, the amplitude of the phasic dopamine burst at US should decrease until a point of convergence, while at the mean time the response at CS should increase until reaching a level higher than the one observed at US. However, one must note, that the fitting of the intermediate group is not as good as for STs or GTs, as it regroups behaviours that range from sign-tracking to goal-tracking, such that this is a weak prediction.

There is the possibility that regularly presenting the magazine or the lever could, without pairing with food, lead to responses that are indistinguishable from CRs. However, ample evidence suggests that the development of a sign-tracking or goal-tracking CR is not due to this pseudoconditioning phenomenon, but rather a result of learned CS-US associations. That is, experience with lever-CS presentations or with food US does not account for the acquisition of lever-CS induced directed responding [22,66]. Nonetheless, it should be noted that the current model cannot distinguish between pseudoconditioning CR-like responses and sign-tracking or goal-tracking behaviours. This would require us to introduce more complex MDPs that embed the ITI and can more clearly distinguish between approach and engagement.

Limitations

The Feature-Model-Free system presented in this article was designed as a proof of concept for the use of factored

representations in computational neuroscience. In its present form it updates the value of one feature (the focused one) at a time, and this is sufficient to account for much of the experimental data. It does not address whether multiple features could be processed in parallel, such that multiple synchronized, but independently computed, signals would update distinct values relative to the attention paid to the associated features. Further experiments should be performed to confirm this hypothesis. Subsequently, using factored representations in the Model-Based system was not necessary to account for the experimental data and the question remains whether explaining some phenomena would require it.

While using factored representations, our approach still relies on the discrete-time state paradigm of classical RL, where updates are made at regular intervals. Although such simplification can explain the set of data considered here, one would need to extend this to continuous time if one would like to also model experimental data where rats take more or less time to initiate actions that can vary in duration [14]. The present model, which does not take timing into consideration, cannot account for the fact that STs and GTs both come to approach their preferred stimuli faster and faster as a function of training nor does it make use of the variations of ITI duration. Our attempt to overcome this limitation using the MDP framework was unsuccessful. Focusing on features, it becomes more tempting to deal with the timing of their presence, a property that is known to be learned and to have some impact on behaviours [61,67–69].

Moreover, in the current model, we did not attempt to account for the conditioned orienting responses (i.e. orientation towards the CS) that both STs and GTs exhibit upon CS presentation [25]. However, we hypothesize that such learned orienting responses could be due to state discrimination mechanisms that are not included in the model, and would be better explained with partial observability and actions dedicated to collect information. This is beyond the scope of the current article, but is of interest for future studies.

As evident by the only partial reproduction of the flupentixol effects on the expression of sign- and goal-tracking behaviours, the model is limited by the use of the softmax action-selection mechanism, which is widely used in computational neuroscience [16–19,32,34–36]. In the model, all actions are equal – there is no action with a specific treatment – and the action-selection mechanism necessarily selects an action at each time step. Any reduction in the value of one action favours the selection of all other actions in proportion to their current associated values. In reality, however, blocking the expression of an action would certainly lead mainly to inactivity rather than necessarily picking the alternative and almost never expressed action. One way of improving the model in this direction could be to replace the classical softmax function by a more realistic model of action selection in the basal ganglia (e.g. [70]). In such a model, no action is performed when no output activity gets above a certain threshold. Humphries et al. [32] have shown that changing the exploration level in a softmax function can be equivalent to changing the level of tonic dopamine in the basal ganglia model of Gurney et al. [70]. Interestingly, in the latter model, reducing the level of tonic dopamine results in difficulty in initiating actions and thus produces lower motor behaviour, as is seen in Parkinsonian patients and as can be seen in rats treated with higher doses of flupentixol [14]. Thus a natural sequel to the current model would be to combine it with a more realistic basal ganglia model for action selection.

We simulated the effect of flupentixol as a reduction of the RPE in the learning processes of Model-Free systems to parallel its blockade of the dopamine receptors. While this is sufficient to account for the pharmacological results previously reported [14], it fails to account for some specific aspects that have more recently emerged. Mainly, it is unable to reproduce the instant decreased engagement observed at the very first trial after post-training local injections of flupentixol [25]. Our current approach requires re-learning to see any impact of flupentixol. A better understanding of the mechanisms that enable instant shifts in motivational values, by shifts in the motivational state [71] or the use of drugs [14,25], might be useful to extend the model on such aspects.

We also tried to model the effect of flupentixol on RPEs with a multiplicative effect, as it would have accounted for an instant impact on behaviour. However, it failed to account for the effects of flupentixol on learning of the sign-tracking CRs, as a multiplicative effect only slowed down learning but did not disrupt it. How to model the impact of flupentixol, and dopamine antagonists or drugs such as cocaine remains an open question (e.g. see [72,73]).

Finally, our work does not currently address the anatomical counterpart of ω at the heart of the model, nor the regions of the brain that would match the current Model-Based system and the Feature-Model-Free system. Numerous studies have already discussed the potential substrates of Model-Based/Model-Free systems in the prefrontal cortex/dorsolateral striatum [74], or the dorsomedial and dorsolateral striatum [33,75–78]. The weighted sum integration may suggest a crossed projection of brains regions favouring sign- and goal-tracking behaviours (Model-Based and Feature-Model-Free systems) into a third one. We postulate there is a difference in strength of "connectivity" between such regions in STs vs GTs [79]. Further, one might hypothesize that the core of the nucleus accumbens contributes to the Feature-Model-Free system. The integration and action selection mechanisms would naturally fit within the basal ganglia, stated to contribute to such functions [32,80–82].

Conclusion

Here we have presented a model that accounts for variations in the form of Pavlovian conditioned approach behaviour seen during autoshaping in rats; that is, the development of a sign-tracking vs goal-tracking CR. This works adds to an emerging set of studies suggesting the presence and collaboration of multiple RL systems in the brain. It questions the classical paradigm of state representation and suggests that further investigation of factored representations in RL models of Pavlovian and instrumental conditioning experiments may be useful.

Methods

Modelling the autoshaping experiment

In the classical reinforcement learning theory [1], tasks are usually described as Markov Decision Processes (MDPs). As the proposed model is based on RL algorithms, we use the MDP formalism to computationally describe the Pavlovian autoshaping procedure used in all simulations.

An MDP describes the interactions of an agent with its environment and the rewards it might receive. An agent being in a state s can execute an action a which results in a new state s' and the possible retrieval of some reward r. More precisely, an agent can be in a finite set of states S, in which it can perform a finite set of discrete actions A, the consequences of which are defined by a transition function $\mathcal{T} : S \times A \rightarrow \Pi(S)$, where $\Pi(S)$ is the probability distribution $\mathcal{P}(s'|s,a)$ of reaching state s' doing action a in state s. Additionally, the reward function $\mathcal{R} : S \times A \rightarrow \mathbb{R}$ is the reward $\mathcal{R}(s,a)$ for doing action a in state s. Importantly, MDPs should theoretically comply with the Markov property: the

probability of reaching state s' should only depend on the last state s and the last action a. An MDP is defined as episodic if it includes at least one state which terminates the current episode.

Figure 1 shows the deterministic MDP used to simulate the autoshaping procedure. Given the variable time schedule (30–150s) and the net difference observed in behaviours in inter-trial intervals, we can reasonably assume that each experimental trial can be simulated with a finite horizon episode.

The agent starts from an empty state (s_0) where there is nothing to do but explore. At some point the lever appears (s_1) and the agent must make a critical choice: It can either go to the lever (s_2) and engage with it (s_5), go to the magazine (s_4) and engage with it (s_7) or just keep exploring (s_3,s_6). At some point, the lever is retracted and food is delivered. If the agent is far from the magazine (s_5,s_7), it first needs to get closer. Once close (s_7), it consumes the food. It ends in an empty state (s_0) which symbolizes the start of the inter-trial interval (ITI): no food, no lever and *an empty but still present magazine*.

The MDP in Figure 1 is common to all of the simulations and independent of the reinforcement learning systems we use. STs should favour the red path, while GTs should favour the *shorter* blue path. All of the results rely mainly on the action taken at the lever appearance (s_1), when choosing to go to either the lever, the magazine, or to explore. Exploring can be understood as not going to the lever nor to the magazine.

To fit with the requirements of the MDP framework, we introduce two limitations in our description, which also simplify our analyses. We assume that engagement is necessarily exclusive to one or no stimulus, and we make no use of the precise timing of the procedure – the ITI duration nor the CS duration – in our simulations.

Inter-trial interval (ITI). While the MDP does not model the ITI, the results regarding physiological data rely partially on its presence. Extending the MDP with a set of states to represent this interval would increase the complexity of the MDP and the time required for simulations. The behaviour that could have resulted from such an extension is easily replaced by applying the following formula at the beginning of each episode:

$$\mathcal{V}(M)\leftarrow(1-u_{ITI})\times\mathcal{V}(M) \tag{1}$$

where the parameter $0\leq u_{ITI}\leq 1$ reflects the interaction with the magazine that occurred during the ITI. A low $u_{ITI}\rightarrow 0$ symbolizes a low interaction and therefore a low revision of the value associated to the magazine. A high $u_{ITI}\rightarrow 1$ symbolizes a strong exploration of the magazine during the inter-trial interval and therefore a strong decrease in the associated value due to unrewarded exploration.

Model

The model relies on the architecture shown in Figure 2. The main idea is to combine the computations of two distinct reinforcement learning systems to define what behavioural response is chosen at each step.

Model-Based system (MB). The first system is Model-Based [1], and classically relies on a transition function $\mathcal{T}$ and a reward function $\mathcal{R}$ which are learned by experience given the following rules:

$$\mathcal{T}(s,a,s')\leftarrow\begin{cases}(1-\alpha)\times\mathcal{T}(s,a,s'')+\alpha & \text{if } s'=s''\\ (1-\alpha)\times\mathcal{T}(s,a,s'') & \text{otherwise}\end{cases} \tag{2}$$

$$\mathcal{R}(s,a)\leftarrow\mathcal{R}(s,a)+\alpha(r-\mathcal{R}(s,a)) \tag{3}$$

where the learning rate $0\leq\alpha\leq 1$ classically represents the speed at which new experiences replace old ones. Using a learning rate rather than counting occurrences is a requirement for accordance with the incremental expression of the observed behaviours. This can account for some resistance or uncertainty in learning from new experiences.

Given this model, an action-value function $\mathcal{Q}$ can then be computed with the following classical formula:

$$\mathcal{Q}(s,a)\leftarrow\mathcal{R}(s,a)+\gamma\sum_{s'}\mathcal{T}(s'|s,a)\max_{a'}\mathcal{Q}(s',a') \tag{4}$$

where the discount rate $0\leq\gamma\leq 1$ classically represents the preference for immediate versus distant rewards. The resulting Advantage function $\mathcal{A}$ [83,84], the output of the first system, is computed as follows:

$$\mathcal{A}(s,a)\leftarrow\mathcal{Q}(s,a)-\max_{a'}\mathcal{Q}(s,a') \tag{5}$$

It defines the (negative) advantage of taking action a in state s relatively to the optimal action known. The optimal action therefore has an advantage value of 0.

In terms of computation, the advantage function could be replaced by the action-value function without changing the simulation results (we only compare $\mathcal{A}-\text{values}$ over the same state and therefore $\max_{a'}\mathcal{Q}(s,a')$ is constant whatever the action). It has been used in preceding works dealing with interactions between instrumental and Pavlovian conditioning [16,84] and we kept it for a better and more straightforward comparison with variants of the model that were directly inspired by these preceding works.

Feature-Model-Free system (FMF). A state is generally described by multiple features. Animals, especially engaged in a repetitive task, might not pay attention to all of them at once. For example, when the lever appears and a rat decides to engage with the magazine, it focuses primarily on the magazine while ignoring the lever, such that it could update a value associated to the magazine but leave intact any value related to the lever (see Figure 10 A). Although this could be related to an attentional process that bias learning, we do not pretend to model attention with such a mechanism.

Relying on this idea, the second system is a revision of classical Model-Free systems which is based on features rather than states. It relies on a value function $\mathcal{V}:\mathcal{C}\rightarrow\mathbb{R}$ based on a set of features $\mathcal{C}$, which is updated with an RPE:

$$\mathcal{V}(c(s,a))\leftarrow\mathcal{V}(c(s,a))+\alpha\delta \tag{6}$$

$$\delta\leftarrow r+\gamma\max_{a'}\mathcal{V}(c(s',a'))-\mathcal{V}(c(s,a))$$

where $c:\mathcal{S}\times\mathcal{A}\rightarrow\mathcal{C}$ is a feature-function that returns the feature $c(s,a)$ the action a was focusing on in state s (see Table S2; Figure 1 also embeds the features returned by c for each action and state). One could argue that this feature-function, defined *a priori*, introduces an additional requirement relative to classical Model-Free systems. This is a weak requirement since this function is straightforward when actions, instead of being abstractly defined,

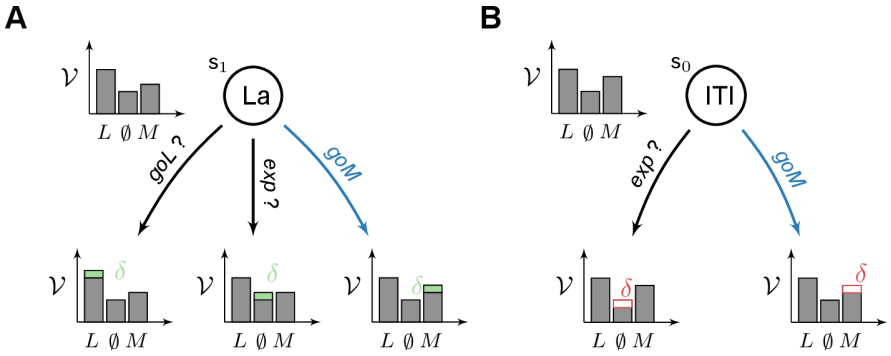


Figure 10. Characteristics of the Feature-Model-Free system. (**A**) Focusing on a particular feature. The Feature-Model-Free system relies on a value function $\mathcal{V}$ based on features. Choosing an action (e.g. *goL*, *goM* or *exp*), defines the feature it is focusing on (e.g. *L*ever, *M*agazine or nothing $\varnothing$). Once the action is chosen (e.g. *goM* in blue), only the value of the focused feature (e.g. $\mathcal{V}(M)$) is updated by a standard reward prediction error, while leaving the values of the other features unchanged. (**B**) Feature-values permit generalization. At a different place and time in the episode, the agent can choose an action (e.g. *goM* in blue) focusing on a feature (e.g. *M*) that might have already been focused on. This leads to the revision of the same value (e.g. $\mathcal{V}(M)$) for two different states (e.g. s_1 and s_0). Values of features are shared amongst multiple states.
doi:10.1371/journal.pcbi.1003466.g010

are described as interactions towards objects in the environment. This function simply states that, for example, when pressing a lever, the animal is focusing on the lever rather than on the magazine. Similar to $\mathcal{Q}-$learning, we assume that the future action to be chosen is the most rewarding one. Therefore, the value chosen for the reached state s', in the computation of the RPE, is the highest value reachable by any possible future action $max_{a'}\mathcal{V}(c(s',a'))$.

Classical Model-Free systems do not permit generalization in their standard form: even when two states share most of their features, updating the value of one state leaves the value of the other untouched. This new system overcomes such limitation (see Figure 10 B). In Feature-Model-Free Reinforcement Learning, multiple states in time and space can share features and their associated values. For example, while in ITI, rats tend from time to time to explore the magazine [22,26], which might lead them to revise any associated value, which can also be used when the lever appears. Therefore, actions in ITIs might impact the rest of the experiment.

In the simulated experiment (see Figure 1), this generalization phenomenon happens as follows: Assuming that the simulated rat was engaging the magazine (eng) before food delivery (from s_4 to s_7), then the value $\mathcal{V}$ of $c(s_4,\text{eng})=M$ is updated with the following $\delta=0+\gamma max_{a'}\mathcal{V}(c(s_7,a'))-\mathcal{V}(M)$. As the best subsequent action (and, for simplification, the only possible one) is to consume the food (in s_7), it results in a positive $\delta=\gamma\mathcal{V}(F)-\mathcal{V}(M)$. During ITI (which in the MDP is simulated by the u_{ITI} parameter), if the simulated rat checks the magazine (goM) and finds no food, then $\mathcal{V}(M)$ is revised with a negative $\delta=\gamma\mathcal{V}(\varnothing)-\mathcal{V}(M)$ (Figure 10 B). The value $\mathcal{V}(M)$ is therefore revised at multiple times in the experiment and, for example, a decrease of value during ITI has an impact on the choice of engaging with the magazine (goM) at lever appearance.

Processing features rather than states and the generalization that results from it is a key mechanism of the presented model. It makes the system favour a different path than the one favoured by classical reinforcement learning systems.

Contrary to what the system suggests, it is almost certain that rats might handle multiple features at once and could simultaneously update multiple values. We present here a version without such capacity since it is not required in the simulated experiments and simplifies its understanding.

Integration. The Feature-Model-Free system accounts for motivational bonuses $\mathcal{V}$ that impact values $\mathcal{A}$ computed by the Model-Based system. The integration of these values is made through a weighted sum:

$$\mathcal{P}(s,a)=(1-\omega)\mathcal{A}(s,a)+\omega\mathcal{V}(c(s,a)) \tag{7}$$

where $0\leq\omega\leq1$ is a combination parameter which defines the importance of each system in the overall model. ω is equivalent to the responsibility signal in Mixture of Experts [35,85]. We want to emphasize that the two systems are not in simple competition, and it is not the case that there is a unique system acting at a time. Rather, they are both active and take part in the decision proportionally to the fixed parameter ω. A simple switch between systems would not account for the full spectrum of observed behaviours ranging from STs to GTs [26].

Action selection. We use a softmax rule on the integrated values $\mathcal{P}$ to compute the probability to select an action A in state s:

$$p(a=A)=\frac{e^{\mathcal{P}(s,A)/\beta}}{\sum_{a'}e^{\mathcal{P}(s,a')/\beta}} \tag{8}$$

where $\beta>0$ is the selection temperature that defines how probabilities are distributed. A high temperature ($\beta\to\infty$) makes all actions equiprobable, a low one makes the most rewarding action almost exclusive.

Impact of flupentixol. When simulating the pharmacological experiments, namely the impact of flupentixol, a parameter $0\leq f<1$ is used to represent the impact of flupentixol on parts of the model.

As a dopamine receptor antagonist, we model the impact of flupentixol on phasic dopamine by revising any RPE δ used in the model given the following formula:

$$\delta_f\leftarrow\begin{cases}\delta-f & \text{if } \dfrac{\delta-f}{\delta}\geq0\\ 0 & \text{otherwise}\end{cases} \tag{9}$$

where δ_f is the new RPE after flupentixol injection. The impact is filtered ($\frac{\delta-f}{\delta}\geq0$) such that flupentixol injection could not lead to negative learning when the RPE was positive, but at most block it

(i.e. the sign of δ_f cannot be different from the one of δ). With a low $f \to 0$, the RPE is not affected ($\delta_f \to \delta$). A high $f \to 1$ reduces the RPE, imitating a blockade of dopamine receptors.

Various studies (e.g. [32]) also suggest that tonic dopamine has an impact on action selection such that any decrease in dopamine level results in favouring exploration over exploitation. We therefore simulated the effect of flupentixol on action selection by revising the selection temperature given the following formula:

$$\beta_f \leftarrow \frac{\beta}{1-f} \tag{10}$$

where β_f is the new selection temperature, and $0 \leq f < 1$ represents the strength of the flupentixol impact. A strong $f \to 1$, which represents an effective dose of flupentixol, favours a high temperature $\beta_f \to \infty$ and therefore exploration. A low $f \to 0$, i.e. a low dose or an absence of flupentixol, leaves the temperature unaffected: $\beta_f \to \beta$.

For the first pharmacological experiment (Effects of systemic flupentixol administration on the learning of sign- and goal-tracking behaviours) both the impact on the softmax and on the RPE were activated, as the flupentixol was injected systemically and assumed to diffuse in the whole brain. For the second experiment (Effects of local flupentixol administration on the expression of sign- and goal-tracking behaviours) only the impact on the RPE was activated, as the flupentixol was injected locally in the core of the nucleus accumbens. We hypothesize that the Feature-Model-Free system relies in the core of the nucleus accumbens whereas the selection process (softmax) does not.

Initialization. In the original experiments [14,20], prior to the autoshaping procedure, rats are familiarized with the Skinner box and the delivery of food into the magazine. While the MDP does not account for such pretraining, we can initialize the model with values ($\mathcal{Q}_i(s_1,goL)$, $\mathcal{Q}_i(s_1,goM)$ and $\mathcal{Q}_i(s_1,exp)$) that reflect it (see the estimation of the model parameters). These initial values can be seen as extra parameters common to the model and its variants.

Variants

Given the modular architecture of the model, we were able to test different combinations of RL systems. Their analysis underlined the key mechanisms required for reproducing each result (see Figures S1, S2, S4 and S5). Figure 11 (B, C and D) schematically represents the analysed variants.

Most of the results rely on the action taken by the agent at the lever appearance. The action taken results from the values $\mathcal{P}(s_1,goL)$, $\mathcal{P}(s_1,goM)$ and $\mathcal{P}(s_1,exp)$, the computation of which differs in each of the variants described below.

Variant 1 : Model-Free/Feature-Model-Free. Variant 1 was tested to assert the necessity of the Model-Based system as part of the model to reproduce the results. Thus in Variant 1, the Model-Based system is replaced by a classical Model-Free system, Advantage learning [83,84], while the Feature-Model-Free system remains unchanged (see Figure 11 B).

In such a Model-Free system, the action-value function $\mathcal{Q}_{\mathrm{MF}}$ is updated online according to the transition just experienced. At each time step the function is updated given an RPE δ that computes the difference between the observed and the expected value, as follows:

$$\mathcal{Q}_{\mathrm{MF}}(s,a) \leftarrow \mathcal{Q}_{\mathrm{MF}}(s,a) + \alpha\delta \tag{11}$$

$$\delta \leftarrow r + \gamma \max_{a'} \mathcal{Q}_{\mathrm{MF}}(s',a') - \mathcal{Q}_{\mathrm{MF}}(s,a)$$

Computation of the associated Advantage function $\mathcal{A}_{\mathrm{MF}}$ follows Equation (5). This model computes integrated values as follows:

$$\mathcal{P}(s,a) = (1-\omega)\mathcal{A}_{\mathrm{MF}}(s,a) + \omega\mathcal{V}(c(s,a)) \tag{12}$$

It is important to note that while Equation (12) looks similar to Equation (7), the Advantage function is computed by a Model-Based system in the model ($\mathcal{A}$) and a Model-Free system in this variant ($\mathcal{A}_{\mathrm{MF}}$), leading to very different results on pharmacological experiments.

Variant 2 : Asymmetrical. Inspired by a work from Dayan et al. [16], Variant 2 combines a classical Advantage learning system [83,84] with some Bias system taking its values directly from the other system (see Figure 11 C). This system computes the integrated values as follows:

$$\mathcal{P}(s,a) = (1-\omega) \times \mathcal{A}_{\mathrm{MF}}(s,a) + \omega \begin{cases} \mathcal{V}(s) & \text{if } a = goL \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

It asymmetrically gives a bonus to the path that should be taken by STs. In slight discrepancy with the original model, it uses the maximum value over action-value function $\mathcal{Q}_{\mathrm{MF}}$ as the value function $\mathcal{V}_{\mathrm{MF}}$ used to compute the advantage function. Hence, there is a single RPE computed at each step.

Variant 3 : Symmetrical. In the same line as Variant 2, Variant 3 symmetrically gives a bonus to both paths using a classical Advantage learning system in combination with a Pavlovian system. This system computes the integrated values as follows:

$$\mathcal{P}(s,a) = \mathcal{A}_{\mathrm{MF}}(s,a) + \begin{cases} \omega\mathcal{V}(s) & \text{if } a = goL \\ (1-\omega)\mathcal{V}(s) & \text{if } a = goM \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

This model does not exactly fit Equation (7) of the general architecture. It is based on 3 systems, where the real competition is between the two bias systems, whereas the Model-Free system is mainly used to compute the values used by the two others (see Figure 11 D). The rest of the architecture is not impacted.

Variant 4 : Model-Based/Model-Free. Variant 4 was developed to confirm the necessity of a feature-based system. It combines two advantage functions computed from a Model-Based ($\mathcal{A}$) and a Model-Free ($\mathcal{A}_{\mathrm{MF}}$) system.

$$\mathcal{P}(s,a) = (1-\omega)\mathcal{A}(s,a) + \omega\mathcal{A}_{\mathrm{MF}}(s,a) \tag{15}$$

While computed differently, both advantage functions will eventually converge to the same optimal values [1] making both systems favouring the same optimal policy. Note that u_{ITI} cannot be used in this variant as there exists no value over the magazine itself. While varying the parameters might slow down learning or make the process more exploratory, this could never lead to sign-tracking as both systems, whatever the weighting, would favour

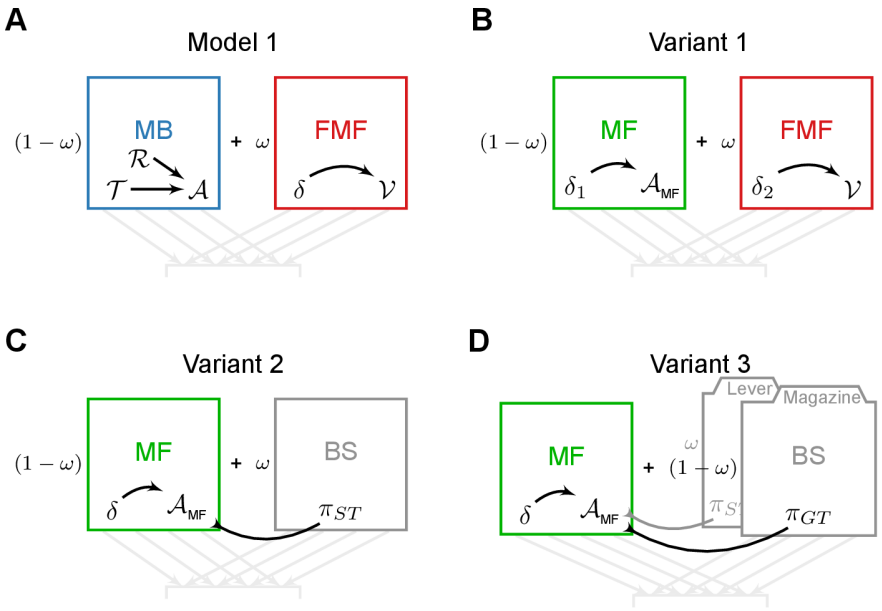


Figure 11. Systems combined in the model and the variants. Variants of the model rely on the same architecture (described in Figure 2) and only differ in the combined systems. Colours are shared for similar systems. (**A**) The model combines a Model-Based system (MB, in blue) and a Feature-Model-Free (FMF, in red) system. (**B**) Variant 1 combines a Model-Free system (MF, in green) and a Feature-Model-Free system. (**C**) Variant 2 combines a Model-Free system and a Bias system (BS, in grey), that relies on values from the Model-Free system. (**D**) Variant 3 combines a Model-Free system and two Bias systems, that rely on values from the Model-Free system. Variant 4 is not included as it failed to even reproduce the autoshaping behavioural results.
doi:10.1371/journal.pcbi.1003466.g011

goal-tracking. As such, Variant 4 is unable to even account for the main behavioural results in the autoshaping procedure (see Figure S8).

Given that all the subsequent simulated results relies on a correct reproduction of the default behaviours, this variant was not investigated further and is not compared to the other variants in supplementary results figures.

Estimating the model parameters

The model relies on model-specific parameters (ω, β, α and γ) and experience-specific parameters (u_{ITI}, $\mathcal{Q}_i(s_1,\text{goL})$, $\mathcal{Q}_i(s_1,\text{goM})$ and $\mathcal{Q}_i(s_1,\varnothing)$). If the model were used to simulate a different experiment, the model-specific parameters would be the same while different experience-specific parameters might be required. For an easier analysis and a simpler comparison between the model and its variants, we reduce the number of parameters by sharing parameters with identical meanings amongst systems (i.e. both systems within the model share values for their learning rates α and discount rates γ, rather than having independent parameter values).

Due to the number of parameters, finding the best values to qualitatively fit the experimental data cannot be done by hand. Using a genetic algorithm makes it possible to optimize the search of suitable values for the parameters.

Parameter values were retrieved by fitting the simulation of the probabilities to engage either the lever or the magazine with the experimental data of one of the previous studies [21]. No direct fitting was intended on other experimental data. Hence, a single set of values was used to simulate behavioural, physiological and pharmacological data.

If for a variant, the optimization algorithm fails to fit the experimental data, it suggests that whatever the values, the mechanisms involved cannot explain the behavioural data (Variant 4).

Probabilities to engage the lever or the magazine were taken as independent objectives of the algorithm, since fitting sign-tracking probabilities is easier than fitting goal-tracking probabilities. For each objective, the fitness function is computed as the least square errors between the experimental and simulated data. Parameter optimization is done with the multi-objective genetic algorithm NSGA-II [86]. We used the implementation provided by the Sferes 2 framework [87]. All parameters required for reproducing the behavioural data were fitted at once.

For NSGA-II, we arbitrarily use a population of 200 individuals and run it over 1000 generations. We use a polynomial mutation with a rate of 0.1, and simulate binary cross-overs with a rate of 0.5. We select the representative individual, to be displayed in figures, from the resulting Pareto front by hand, such that it best visually fits the observed data.

To confirm that ω is the key parameter of the model, we additionally tried to fit the whole population at once (i.e. sharing all parameter values in agents but ω) and we were still able to reproduce the observed tendencies of sign- and goal-tracking in the population (see Figure S7 A,B) and the resulting different phasic dopaminergic patterns (see Figure S7 C,D).

It is however almost certain that each subgroup does not express the exact same values for the other parameters. Removing such constraint by fitting each subgroup separately, indeed provides better results. Results presented in this article are based on such separate fitting.

Supporting Information

Figure S1 Comparison of variants of the model on simulations of autoshaping experiment. Legend is as in Figure 5 (C,D). Simulation parameters for STs (red), GTs (blue)

and IGs (white) in the model (**A**), Variant 1 (**B**), Variant 2 (**C**) and Variant 3 (**D**) are summarized in Table S1. All variants reproduce the spectrum of behaviours ranging from sign-tracking to goal-tracking.
(TIFF)

Figure S2 Comparison of variants of the model on incentive salience and Conditioned Reinforcement Effect intuitions. Legend is as in Figure 6. Simulation parameters for STs (red), GTs (blue) and IGs (white) are summarized in Table S1. Variant 2 (**C**) relying on asymmetrical bonuses given only to sign-tracking cannot reproduce the attribution of a motivational value by the second system to both the lever and the magazine. Others (**A,B,D**) attribute values to both stimuli and parallels the supposed acquisition of motivational values by stimuli, i.e. incentive salience. All variants are able to account for a Conditioned Reinforcement Effect more pronounced in STs than in GTs.
(TIFF)

Figure S3 Comparison of variants of the model on simulations of patterns of dopaminergic activity. Legend is as in Figure 7 (C,D). Simulation parameters for STs (left) and GTs (right) are summarized in Table S1. The model (**A**) and Variant 1 (**B**) can reproduce the difference observed in dopaminergic patterns of activity in STs versus GTs. Other variants (**C,D**) fail to do so, given that the classical Model-Free system propagates the RPE from food delivery to lever appearance on all pathways of the MDP.
(TIFF)

Figure S4 Comparison of variants on simulations of the effect of systemic injections of flupentixol. Legend is as in Figure 8 (C,D). Simulation parameters for STs (left) and GTs (right) are summarized in Table S1. Only the Model (**A**) can reproduce the difference in response to injections of flupentixol observed in STs versus GTs. All variants (**B,C,D**) fail to do so, given that they only rely on Model-Free, i.e. RPE-dependent, mechanisms that are blocked by flupentixol.
(TIFF)

Figure S5 Comparison of variants on simulations of the effect of post injections of flupentixol. Legend is as in Figure 9 (C,D). Simulation parameters for groups of rats composing the population are summarized in Table S1. Variants 2 (**C**) and 3 (**D**), accounting for sign- and goal-tracking using a single set of values, have a similar impact of flupentixol on both behaviours, leaving relative probabilities to engage with lever and magazine unaffected. Variant 1 (**B**) uses different systems, thus flupentixol impacts sign-tracking in the model in the same way as it does in experimental data. However, given that both systems rely on RPE-dependent mechanisms, the impact is not as visible as in the model (**A**).
(TIFF)

Figure S6 Prediction of the model about expected patterns of dopaminergic activity in intermediate groups. Data are expressed as mean $\pm$ S.E.M. Average RPE computed by the Feature-Model-Free system in response to CS and US presentation for each session of conditioning in the intermediate group. Simulated group is defined as in Figure 5.
(TIFF)

Figure S7 Behavioural and physiological simulations of autoshaping with shared parameter values across STs, GTs and IGs. (**A,B**) Legend is as in Figure 5 (C,D). Reproduction of the respective tendencies to sign- and goal-track of STs ($\omega = 0.5$), IGs ($\omega = 0.375$) and GTs ($\omega = 0.05$)) using a single set of parameters ($\alpha = 0.2$, $\gamma = 0.8$, $\beta = 0.09$, $u_{ITI} = 0.2$, $\mathcal{Q}_i(s_1, goL) = 0.0$, $\mathcal{Q}_i(s_1, exp) = 0.5$ and $\mathcal{Q}_i(s_1, goM) = 0.5$). (**C,D**) Legend is as in Figure 7 (C,D). Reproduction of the different patterns of phasic dopaminergic activity in STs and GTs using the same single set of parameters. By simply varying the ω parameter, the model can still qualitatively reproduce the observations in experimental data.
(TIFF)

Figure S8 Simulation of autoshaping experiment for Variant 4. Legend is as in Figure 5 (C,D). Simulation for parameters STs (red), GTs (blue) and IGs (white) in the Variant 4 are summarized in Table S1. Variant 4 is not even able to reproduce the main behavioural data.
(TIFF)

Table S1 Summary of parameters used in simulations. Parameters retrieved by optimisation with NSGA-II and used to produce the results presented in this article for the model and its variants. Parameters for STs, GTs and IGs were optimized separately (A,B,C,D,E). To confirm that ω is the key parameter of the model, we also optimized parameters for STs, GTs and IGs by sharing all but the ω parameter (F) to produce Figure S7.
(TIFF)

Table S2 Definition of feature-function c. Stimuli (*L*ever, *M*agazine, *F*ood or $\varnothing$) returned by the feature-function c for each possible state-action pair $\langle s,a \rangle$ in the MDP described in Figure 1. The feature-function simply defines the stimulus that is the focus of an action in a particular state.
(TIFF)

Acknowledgments

The authors would like to thank Angelo Arleo, Kent Berridge, Etienne Coutureau, Alain Marchand and Benjamin Saunders for helpful discussions. The authors would also like to thank the reviewers for their valuable comments and suggestions that helped to improve the contents of this paper.

Author Contributions

Conceived and designed the experiments: FL OS SBF TER MK. Performed the experiments: FL. Analyzed the data: FL OS SBF TER MK. Contributed reagents/materials/analysis tools: FL SBF TER. Wrote the paper: FL OS SBF TER MK.

References

1. Sutton RS, Barto AG (1998) Reinforcement learning: An introduction. The MIT Press.
2. Sutton RS, Barto AG (1987) A temporal-difference model of classical conditioning. In: Proceedings of the ninth annual conference of the cognitive science society. Seattle, WA, pp. 355–378.
3. Barto AG (1995) Adaptive critics and the basal ganglia. In: Houk JC, Davis JL, Beiser DG, editors, Models of information processing in the basal ganglia, The MIT Press. pp. 215–232.
4. Clark JJ, Hollon NG, Phillips PEM (2012) Pavlovian valuation systems in learning and decision making. Curr Opin Neurobiol 22: 1054–1061.
5. Simon DA, Daw ND (2012) Dual-system learning models and drugs of abuse. In: Computational Neuroscience of Drug Addiction, Springer. pp. 145–161.
6. Cardinal RN, Parkinson JA, Hall J, Everitt BJ (2002) Emotion and motivation: the role of the amygdala, ventral striatum, and prefrontal cortex. Neurosci Biobehav Rev 26: 321–352.

7. Yin HH, Ostlund SB, Knowlton BJ, Balleine BW (2005) The role of the dorsomedial striatum in instrumental conditioning. Eur J neurosci 22: 513–523.
8. Solway A, Botvinick MM (2012) Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates. Psychol Rev 119: 120–154.
9. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ (2011) Model-based influences on humans' choices and striatal prediction errors. Neuron 69: 1204–1215.
10. Graybiel AM (2008) Habits, rituals, and the evaluative brain. Annu Rev Neurosci 31: 359–387.
11. Yin HH, Knowlton BJ, Balleine BW (2004) Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. Eur J neurosci 19: 181–189.
12. Schultz W (1998) Predictive reward signal of dopamine neurons. J Neurophysiol 80: 1–27.
13. Fiorillo CD, Tobler PN, Schultz W (2003) Discrete coding of reward probability and uncertainty by dopamine neurons. Science 299: 1898–1902.
14. Flagel SB, Clark JJ, Robinson TE, Mayo L, Czuj A, et al. (2011) A selective role for dopamine in stimulus-reward learning. Nature 469: 53–57.
15. Danna CL, Elmer GI (2010) Disruption of conditioned reward association by typical and atypical antipsychotics. Pharmacol Biochem Behav 96: 40–47.
16. Dayan P, Niv Y, Seymour B, Daw ND (2006) The misbehavior of value and the discipline of the will. Neural Netw 19: 1153–1160.
17. Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci 8: 1704–1711.
18. Keramati M, Dezfouli A, Piray P (2011) Speed/Accuracy trade-off between the habitual and the goal-directed processes. PLoS Comput Biol 7: e1002055.
19. Gläscher J, Daw ND, Dayan P, O'Doherty JP (2010) States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron 66: 585–595.
20. Flagel SB, Watson SJ, Robinson TE, Akil H (2007) Individual differences in the propensity to approach signals vs goals promote different adaptations in the dopamine system of rats. Psychopharmacology 191: 599–607.
21. Flagel SB, Akil H, Robinson TE (2009) Individual differences in the attribution of incentive salience to reward-related cues: Implications for addiction. Neuropharmacology 56: 139–148.
22. Robinson TE, Flagel SB (2009) Dissociating the predictive and incentive motivational properties of reward-related cues through the study of individual differences. Biol psychiatry 65: 869–873.
23. Mahler SV, Berridge KC (2009)Which cue to "want?" Central amygdala opioid activation enhances and focuses incentive salience on a prepotent reward cue. J Neurosci 29: 6500–13.
24. DiFeliceantonio AG, Berridge KC (2012) Which cue to 'want'? Opioid stimulation of central amygdala makes goal-trackers show stronger goal-tracking, just as sign-trackers show stronger sign-tracking. Behav Brain Res 230: 399–408.
25. Saunders BT, Robinson TE (2012) The role of dopamine in the accumbens core in the expression of pavlovian-conditioned responses. Eur J neurosci 36: 2521–2532.
26. Meyer PJ, Lovic V, Saunders BT, Yager LM, Flagel SB, et al. (2012) Quantifying individual variation in the propensity to attribute incentive salience to reward cues. PLoS ONE 7: e38987.
27. Berridge KC (2007) The debate over dopamines role in reward: the case for incentive salience. Psychopharmacology 191: 391–431.
28. Lovic V, Saunders BT, Yager LM, Robinson TE (2011) Rats prone to attribute incentive salience to reward cues are also prone to impulsive action. Behav Brain Res 223: 255–261.
29. Williams BA (1994) Conditioned reinforcement: Experimental and theoretical issues. Behav Anal 17: 261–285.
30. Skinner BF (1938) The behavior of organisms: An experimental analysis. Appleton-Century-Crofts New York, 82–82 pp.
31. Lomanowska AM, Lovic V, Rankine MJ, Mooney SJ, Robinson TE, et al. (2011) Inadequate early social experience increases the incentive salience of reward-related cues in adulthood. Behav Brain Res 220: 91–99.
32. Humphries MD, Khamassi M, Gurney K (2012) Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. Front Neurosci 6: 9.
33. Khamassi M, Humphries MD (2012) Integrating cortico-limbic-basal ganglia architectures for learning model-based and model-free navigation strategies. Front Behav Neurosci 6.
34. Huys QJM, Eshel N, O'Nions E, Sheridan L, Dayan P, et al. (2012) Bonsai trees in your head: How the pavlovian system sculpts goal-directed choices by pruning decision trees. PLoS Comput Biol 8: e1002410.
35. Doya K, Samejima K, Katagiri Ki, Kawato M (2002) Multiple model-based reinforcement learning. Neural Comput 14: 1347–1369.
36. Redish AD, Jensen S, Johnson A, Kurth-Nelson Z (2007) Reconciling reinforcement learning models with behavioral extinction and renewal: Implications for addiction, relapse, and problem gambling. Psychol Rev 114: 784–805.
37. Takahashi YK, Roesch MR, Stalnaker TA, Haney RZ, Calu DJ, et al. (2009) The orbitofrontal cortex and ventral tegmental area are necessary for learning from unexpected outcomes. Neuron 62: 269–280.
38. McDannald MA, Lucantonio F, Burke KA, Niv Y, Schoenbaum G (2011) Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. J Neurosci 31: 2700–2705.
39. McDannald MA, Takahashi YK, Lopatina N, Pietras BW, Jones JL, et al. (2012) Model-based learning and the contribution of the orbitofrontal cortex to the model-free world. Eur J neurosci 35: 991–996.
40. Cleland GG, Davey GCL (1983) Autoshaping in the rat: The effects of localizable visual and auditory signals for food. J Exp Anal Behav 40: 47–56.
41. Meyer PJ, Aldridge JW, Robinson TE (2010) Auditory and visual cues are differentially attributed with incentive salience but similarly affected by amphetamine, 2010 neuroscience meeting planner. In: Society for Neuroscience Annual Meeting (SfN10).
42. Schmajuk NA, Lam YW, Gray JA (1996) Latent inhibition: A neural network approach. J Exp Psychol Anim Behav Process 22: 321–349.
43. Balkenius C (1999) Dynamics of a classical conditioning model. Auton Robots 7: 41–56.
44. Stout SC, Miller RR (2007) Sometimes-competing retrieval (SOCR): A formalization of the comparator hypothesis. Psychol Rev 114: 759–783.
45. Courville AC, Daw ND, Touretzky DS (2006) Bayesian theories of conditioning in a changing world. Trends Cogn Sci 10: 294–300.
46. Gershman SJ, Niv Y (2012) Exploring a latent cause theory of classical conditioning. Anim Learn Behav 40: 255–268.
47. Kamin LJ (1967) Predictability, surprise, attention, and conditioning. In: Campbell BA, Church RMa, editors, Punishment and aversive behavior, New York: Appleton-Century-Crofts. pp. 279–296.
48. Lattal KM, Nakajima S (1998) Overexpectation in appetitive pavlovian and instrumental conditioning. Anim Learn Behav 26: 351–360.
49. Bellman R (1957) Dynamic programming. Princeton University Press.
50. Khamassi M, Martinet LE, Guillot A (2006) Combining self-organizing maps with mixtures of experts: application to an actor-critic model of reinforcement learning in the basal ganglia. In: From Animals to Animats 9, Springer. pp. 394–405.
51. Elfwing S, Uchibe E, Doya K (2013) Scaled free-energy based reinforcement learning for robust and efficient learning in high-dimensional state spaces. Front Neurorobot 7: 3.
52. Boutilier C, Dearden R, Goldszmidt M (2000) Stochastic dynamic programming with factored representations. Artif Intell 121: 49–107.
53. Degris T, Sigaud O, Wuillemin PH (2006) Learning the structure of factored markov decision processes in reinforcement learning problems. In: Proceedings of the 23rd international conference on Machine learning. ACM, pp. 257–264.
54. Vigorito CM, Barto AG (2008) Autonomous hierarchical skill acquisition in factored mdps. In: Yale Workshop on Adaptive and Learning Systems, New Haven, Connecticut. volume 63, p. 109.
55. Guitart-Masip M, Huys QJM, Fuentemilla L, Dayan P, Duzel E, et al. (2012) Go and no-go learning in reward and punishment: interactions between affect and effect. Neuroimage 62: 154–166.
56. Huys QJM, Cools R, Gölzer M, Friedel E, Heinz A, et al. (2011) Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. PLoS Comput Biol 7: e1002028.
57. Yin HH, Ostlund SB, Balleine BW (2008) Reward-guided learning beyond dopamine in the nucleus accumbens: the integrative functions of cortico-basal ganglia networks. Eur J neurosci 28: 1437–1448.
58. Corbit LH, Balleine BW (2005) Double dissociation of basolateral and central amygdala lesions on the general and outcome-specific forms of pavlovian-instrumental transfer. J Neurosci 25: 962–970.
59. Balsam PD, Payne D (1979) Intertrial interval and unconditioned stimulus durations in autoshaping. Anim Learn Behav 7: 477–482.
60. Gibbon J, Balsam P (1981) Spreading association in time, Academic Press. pp. 219–253.
61. Gallistel CR, Gibbon J (2000) Time, rate, and conditioning. Psychol Rev 107: 289–344.
62. Tomie A, Festa ED, Sparta DR, Pohorecky LA (2003) Lever conditioned stimulus–directed autoshaping induced by saccharin–ethanol unconditioned stimulus solution: effects of ethanol concentration and trial spacing. Alcohol 30: 35–44.
63. Morris G, Nevet A, Arkadir D, Vaadia E, Bergman H (2006) Midbrain dopamine neurons encode decisions for future action. Nat Neurosci 9: 1057–1063.
64. Roesch MR, Calu DJ, Schoenbaum G (2007) Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. Nat Neurosci 10: 1615–1624.
65. Bellot J, Sigaud O, Khamassi M (2012) Which temporal difference learning algorithm best reproduces dopamine activity in a multi-choice task? In: From Animals to Animats 12, Springer. pp. 289–298.
66. Tomie A, Lincks M, Nadarajah SD, Pohorecky LA, Yu L (2012) Pairings of lever and food induce pavlovian conditioned approach of sign-tracking and goal-tracking in c57bl/6 mice. Behav Brain Res 226: 571–578.
67. Kobayashi S, Schultz W (2008) Influence of reward delays on responses of dopamine neurons. J Neurosci 28: 7837–7846.
68. Daw ND, Courville AC, Touretzky DS (2006) Representation and timing in theories of the dopamine system. Neural Comput 18: 1637–1677.
69. Fiorillo CD, Newsome WT, Schultz W (2008) The temporal precision of reward prediction in dopamine neurons. Nat Neurosci 11: 966–973.
70. Gurney KN, Humphries MD, Wood R, Prescott TJ, Redgrave P (2004) Testing computational hypotheses of brain systems function: a case study with the basal ganglia. Network 15: 263–290.

71. Robinson MJF, Berridge KC (2013) Instant transformation of learned repulsion into motivational "wanting". Current Biology 23: 282–289.
72. Panlilio LV, Thorndike EB, Schindler CW (2007) Blocking of conditioning to a cocaine-paired stimulus: testing the hypothesis that cocaine perpetually produces a signal of larger-than-expected reward. Pharmacol Biochem Behav 86: 774–777.
73. Redish AD (2004) Addiction as a computational process gone awry. Science 306: 1944–1947.
74. Daw ND, Niv Y, Dayan P (2006) Actions, policies, values and the basal ganglia. In: Bezard E, editor, Recent Breakthroughs in Basal Ganglia Research, Nova Science Publishers, Inc Hauppauge, NY. pp. 91–106.
75. Yin HH, Knowlton BJ (2006) The role of the basal ganglia in habit formation. Nat Rev Neurosci 7: 464–476.
76. Thorn CA, Atallah H, Howe M, Graybiel AM (2010) Differential dynamics of activity changes in dorsolateral and dorsomedial striatal loops during learning. Neuron 66: 781–795.
77. Bornstein AM, Daw ND (2011) Multiplicity of control in the basal ganglia: computational roles of striatal subregions. Curr Opin Neurobiol 21: 374–380.
78. van der Meer M, Kurth-Nelson Z, Redish AD (2012) Information processing in decision-making systems. Neuroscientist 18: 342–359.
79. Flagel SB, Cameron CM, Pickup KN, Watson SJ, Akil H, et al. (2011) A food predictive cue must be attributed with incentive salience for it to induce c-fos mRNA expression in cortico-striatalthalamic brain regions. Neuroscience 196: 80–96.
80. Mink JW (1996) The basal ganglia: focused selection and inhibition of competing motor programs. Prog Neurobiol 50: 381–425.
81. Redgrave P, Prescott TJ, Gurney K (1999) The basal ganglia: a vertebrate solution to the selection problem? Neuroscience 89: 1009–1023.
82. Gurney K, Prescott TJ, Redgrave P (2001) A computational model of action selection in the basal ganglia. I. A new functional anatomy. Biol Cybern 84: 401–410.
83. Baird III LC (1993) Advantage updating. Technical report, DTIC Document.
84. Dayan P, Balleine BW (2002) Reward, motivation, and reinforcement learning. Neuron 36: 285–298.
85. Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE (1991) Adaptive mixtures of local experts. Neural Comput 3: 79–87.
86. Deb K, Pratap A, Agarwal S, Meyarivan T (2002) A fast and elitist multiobjective genetic algorithm: Nsga-ii. IEEE Trans Evol Comput 6: 182–197.
87. Mouret JB, Doncieux S (2010) SFERESv2: Evolvin' in the Multi-Core World. In: WCCI 2010 IEEE World Congress on Computational Intelligence, Congress on Evolutionary Computation (CEC). pp. 4079–4086.

Supporting Information Legends

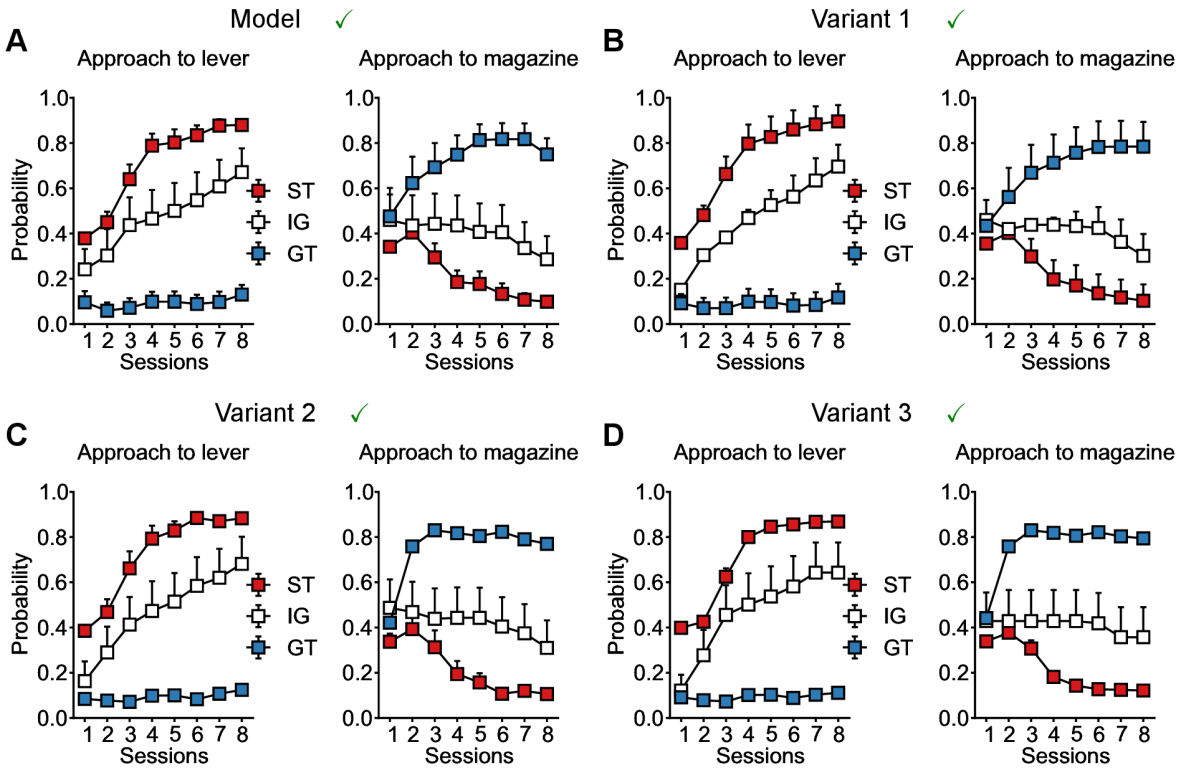


Figure S1. Comparison of variants of the model on simulations of autoshaping experiment. Legend is as in Figure 5 (C,D). Simulation parameters for STs (red), GTs (blue) and IGs (white) in the Model **(A)**, Variant 1 **(B)**, Variant 2 **(C)** and Variant 3 **(D)** are summarized in Table S1. All variants reproduce the spectrum of behaviours ranging from sign-tracking to goal-tracking.

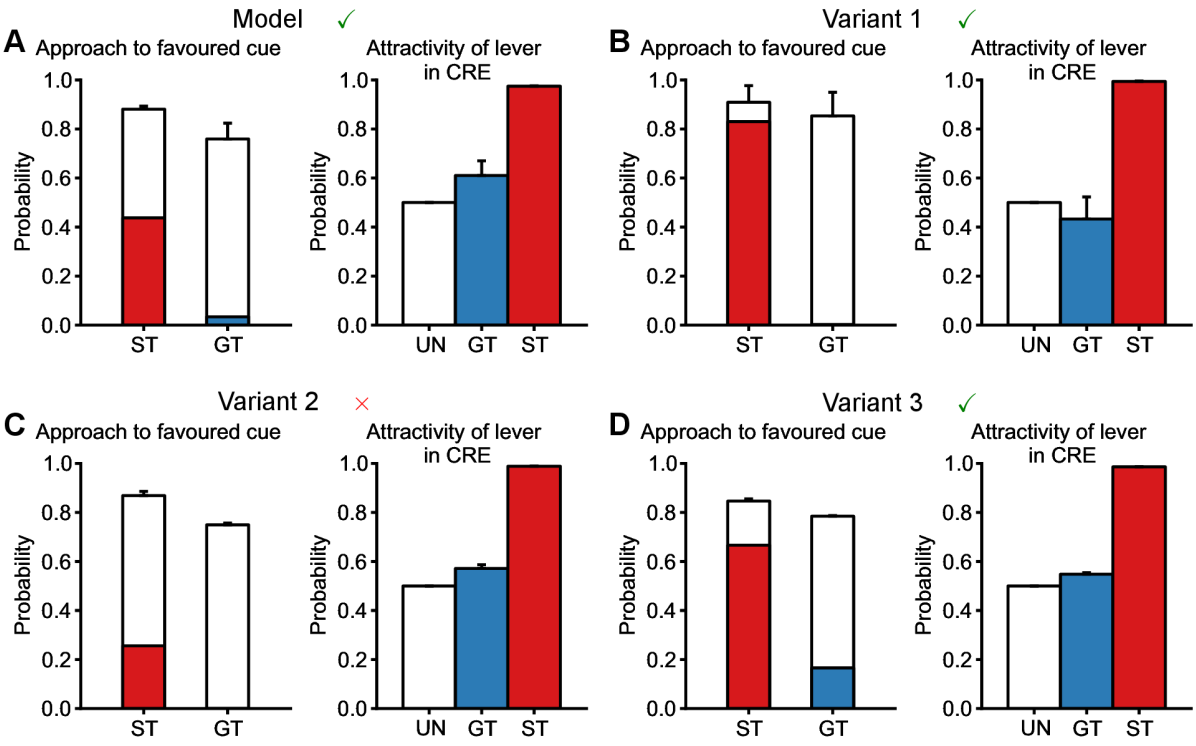


Figure S2. Comparison of variants of the model on incentive salience and Conditioned Reinforcement Effect intuitions. Legend is as in Figure 6. Simulation parameters for STs (red), GTs (blue) and IGs (white) are summarized in Table S1. Variant 2 **(C)** relying on asymmetrical bonuses given only to sign-tracking cannot reproduce the attribution of a motivational value by the second system to both the lever and the magazine. Others **(A,B,D)** attribute values to both stimuli and parallels the supposed acquisition of motivational values by stimuli, i.e. incentive salience. All variants are able to account for a Conditioned Reinforcement Effect more pronounced in STs than in GTs.

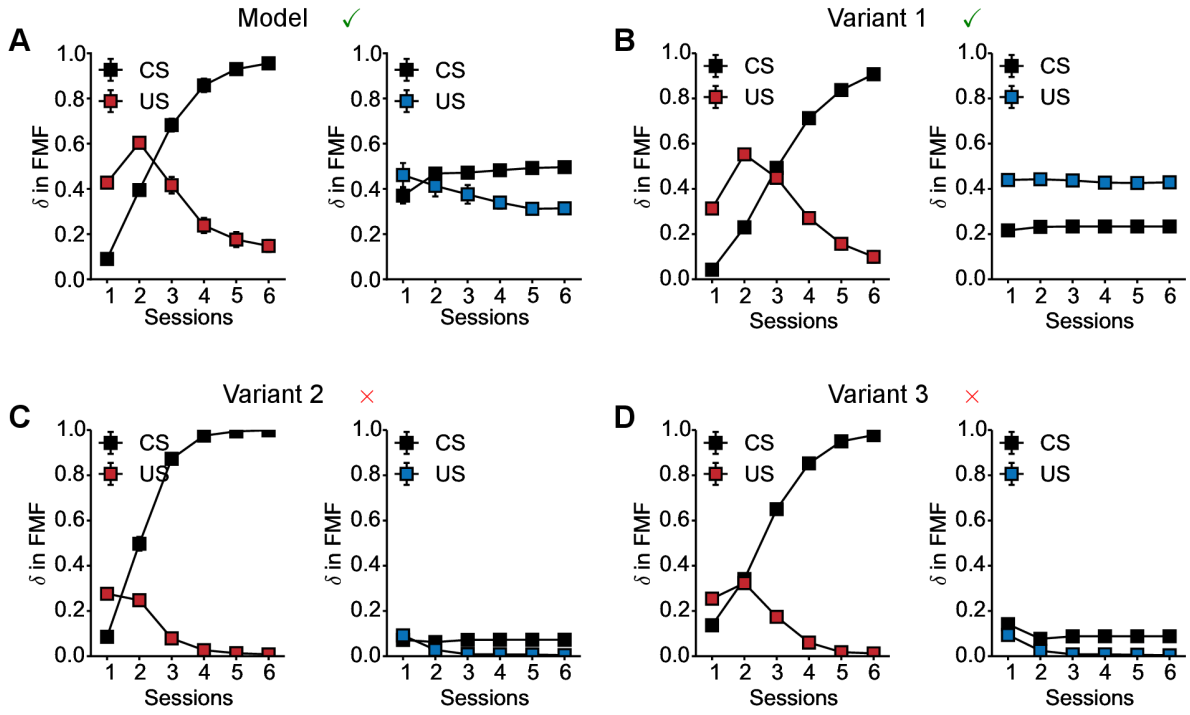


Figure S3. Comparison of variants of the model on simulations of patterns of dopaminergic activity. Legend is as in Figure 7 (C,D). Simulation parameters for STs (left) and GTs (right) are summarized in Table S1. The Model **(A)** and Variant 1 **(B)** can reproduce the difference observed in dopaminergic patterns of activity in STs versus GTs. Other variants **(C,D)** fail to do so, given that the classical Model-Free system propagates the RPE from food delivery to lever appearance on all pathways of the MDP.

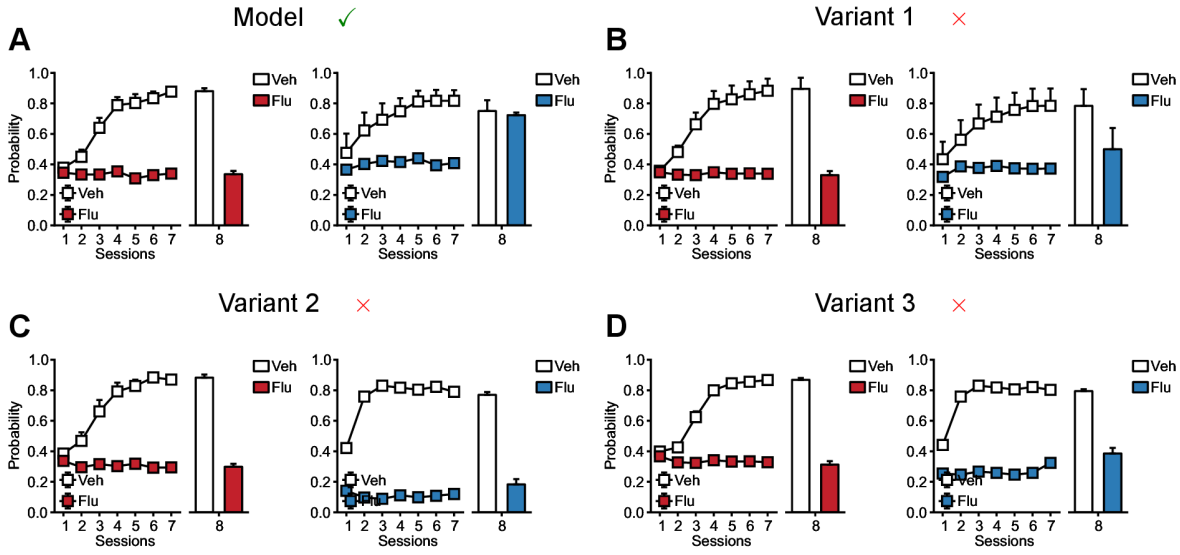


Figure S4. Comparison of variants on simulations of the effect of systemic injections of flupentixol. Legend is as in Figure 8 (C,D). Simulation parameters for STs (left) and GTs (right) are summarized in Table S1. Only the Model **(A)** can reproduce the difference in response to injections of flupentixol observed in STs versus GTs. All variants **(B,C,D)** fail to do so, given that they only rely on Model-Free, i.e. RPE-dependent, mechanisms that are blocked by flupentixol.

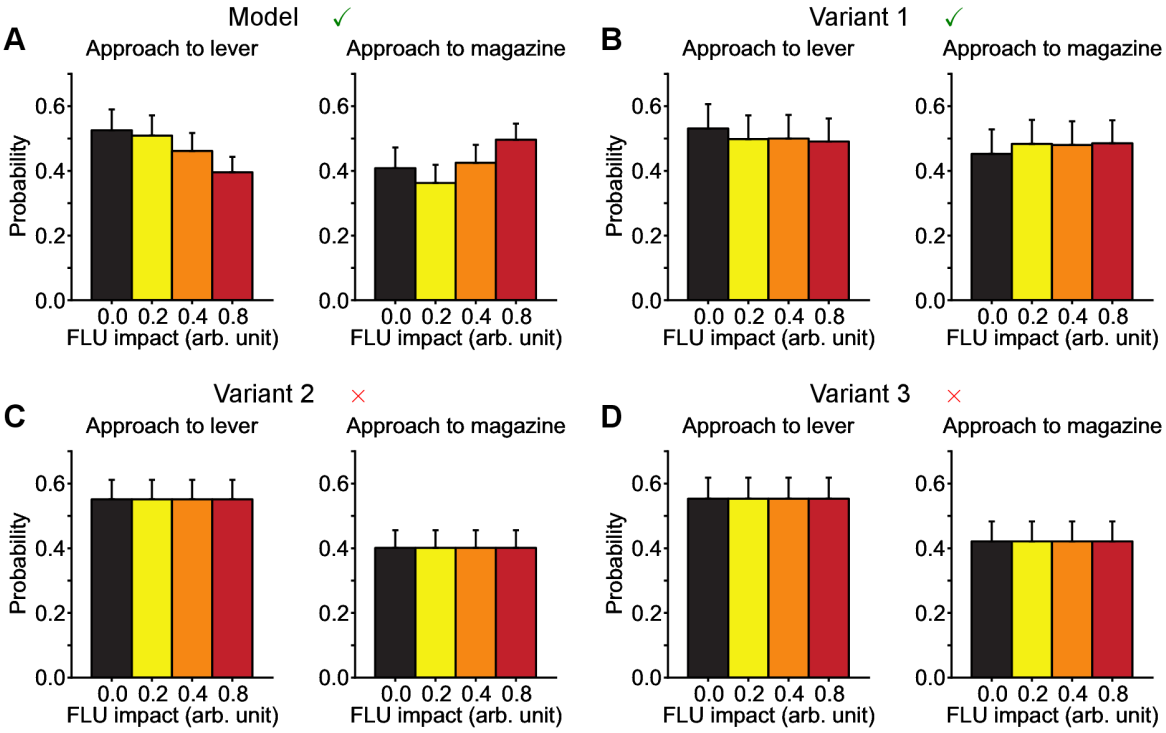


Figure S5. Comparison of variants on simulations of the effect of post injections of flupentixol. Legend is as in Figure 9 (C,D). Simulation parameters for groups of rats composing the population are summarized in Table S1. Variants 2 **(C)** and 3 **(D)** accounting for sign- and goal-tracking using a single set of values have a similar impact of flupentixol on both behaviours leaving relative probabilities to engage with lever and magazine unaffected. Variant 1 **(B)** uses different systems, thus flupentixol impacts sign-tracking in the model in the same way as it does in experimental data. However, given that both systems rely on RPE-dependent mechanisms, the impact is not as visible as in Model 1 **(A)**.

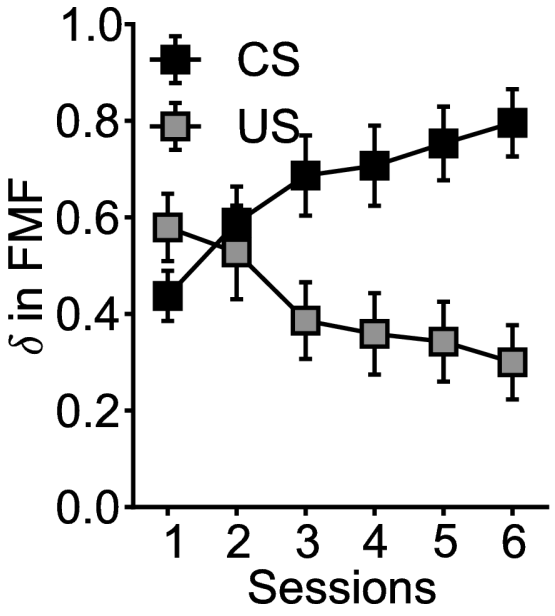


Figure S6. Prediction of the Model about expected patterns of dopaminergic activity in intermediate groups. Data are expressed as mean ± S.E.M. Average RPE computed by the Feature-Model-Free system in response to CS and US presentation for each session of conditioning in the intermediate group. Simulated group is defined as in Figure 5.

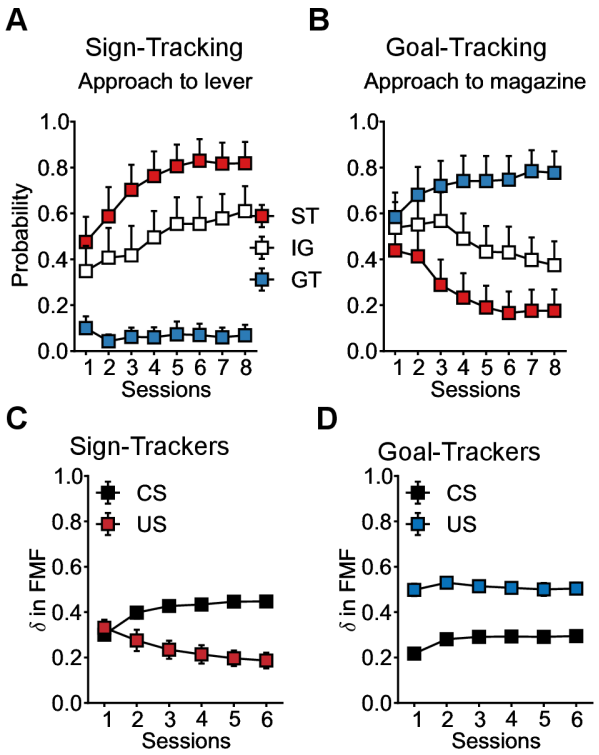


Figure S7. Behavioural and Physiological simulations of autoshaping with shared parameter values across STs, GTs and IGs. (A,B) Legend is as in Figure 5 (C,D). Reproduction of the respective tendencies to sign- and goal-track of STs ($\omega = 0.5$), IGs ($\omega = 0.375$) and GTs ($\omega = 0.05$)) using a single set of parameters ($\alpha = 0.2$, $\gamma = 0.8$, $\beta = 0.09$, $u_{ITI} = 0.2$, $\mathcal{Q}_i(s_1, goL) = 0.0$, $\mathcal{Q}_i(s_1, exp) = 0.5$ and $\mathcal{Q}_i(s_1, goM) = 0.5$). **(C,D)** Legend is as in Figure 7 (C,D). Reproduction of the different patterns of phasic dopaminergic activity in STs and GTs using the same single set of parameters. By simply varying the ω parameters, the model can still qualitatively reproduce the observations in experimental data.

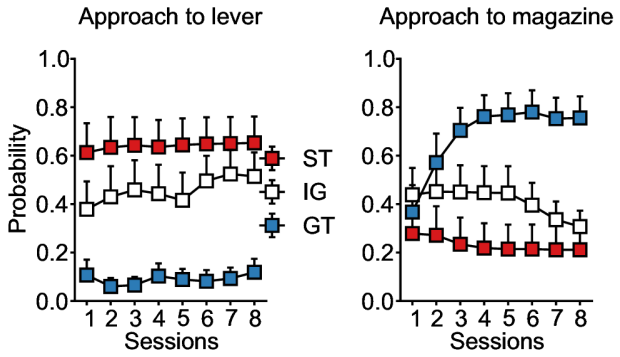


Figure S8. Simulation of autoshaping experiment for Variant 4. Legend is as in Figure 5 (C,D). Simulation for parameters STs (red), GTs (blue) and IGs (white) in the variant 4 are summarized in Table S1. Variant 4 is not even able to reproduce the main behavioural data.

Table S1. Summary of parameters used in simulations

Version	Type	ω	β	α	γ	u_{ITI}	$\mathcal{Q}_i(s_1, L)$	$\mathcal{Q}_i(s_1, \emptyset)$	$\mathcal{Q}_i(s_1, M)$
A: Model	ST	0.499	0.239	0.031	0.996	0.027	0.844	0.999	0.538
	IG	0.276	0.142	0.217	0.999	0.228	0.526	0.888	0.587
	GT	0.048	0.084	0.895	0.727	0.140	1.0	0.316	0.023
B: Variant 1	ST	0.994	0.145	0.018	0.999	0.995	0.278	0.999	0.676
	IG	0.350	0.095	0.023	0.971	0.904	0.398	0.675	0.712
	GT	0.003	0.002	0.906	0.508	0.263	0.147	0.419	0.520
C: Variant 2	ST	0.788	0.367	0.055	0.996	0	0.153	0.133	0.151
	IG	0.843	0.046	0.779	0.999	0	0	0.532	0.593
	GT	0.211	0.130	0.109	0.445	0	0	1	0.095
D: Variant 3	ST	0.295	0.189	0.070	0.999	0	0.057	0.054	0
	IG	0.333	0.027	0.926	0.674	0	0.011	0.444	0.747
	GT	0.166	0.047	0.093	0.417	0	0	0.476	0.229
E: Variant 4	ST	0.643	0.136	0.763	1	-	0.325	0.713	0.094
	IG	0.277	0.175	0.748	0.999	-	0.273	0.784	0.986
	GT	0.529	0.077	0.617	0.695	-	0.102	0.635	0.962
F: Model	ST	0.500	0.090	0.20	0.800	0.200	0.000	0.400	0.400
(shared)	IG	0.375	0.090	0.20	0.800	0.200	0.000	0.400	0.400
	GT	0.050	0.090	0.20	0.800	0.200	0.000	0.400	0.400

Parameters retrieved by optimisation with NSGA-II and used to produce the results presented in this article for the model and its variants. Parameters for STs, GTs and IGs were optimized separately (A,B,C,D,E). To confirm that ω is the key parameter of the model, we also optimized parameters for STs, GTs and IGs by sharing all but the ω parameter (F) to produce Figure S7.

Table S2. Definition of feature-function c

s	s_0	s_1	s_1	s_1	s_2	s_3	s_4	s_5	s_6	s_7
a	exp	goL	exp	goM	eng	$\emptyset$	eng	goM	goM	eat
$c(s,a)$	$\emptyset$	L	$\emptyset$	M	L	$\emptyset$	M	F	F	F

Stimuli (Lever, Magazine, Food or $\emptyset$) returned by the feature-function c for each possible state-action pair $\langle s, a\rangle$ in the MDP described in Figure 1. The feature-function simply defines the stimulus that is the focus of an action in a particular state.

3 Model-based analyses of behavioral and neural correlates of behavioral adaptation

Contents

This chapter presents work employing the model-based analysis of neurophysiological data approach. The work is presented under the form of two journal papers, one in press (Khamassi et al. 2014), the other about to be submitted (Bellot et al. in preparation), aiming at testing model predictions about hypothesized neural activities underlying behavioral adaptation, and using the computational models to more precisely measure information related to particular computational mechanisms in neural activity.

The first one has been performed with Emmanuel Procyk, Peter F. Dominey, René Quilodran and Pierre Enel and shows neural substrates of adaptive regulation of reinforcement learning parameters in the prefrontal cortical network during monkey behavioral adaptation. The results show differences in activity response patterns between the Anterior Cingulate Cortex (ACC) and Lateral Prefrontal Cortex (LPFC) suggesting a role of ACC in integrating reinforcement-based information to regulate decision functions in LPFC under varying control levels, which could be interpreted in terms of varying levels of the exploration parameter in the reinforcement learning model.

The second one presents the work of PhD student Jean Bellot and shows model-based analyses of dopamine neurons' single-unit recordings

during a decision-making task in rats. The work shows that in contrast to previous reports, dopamine activity in this task only partially reflects the computation of a reward prediction error and also incorporates information about the value function. Moreover, the dynamics of this signal appears to be partly disconnected from the dynamics of observed behavioral adaptation, suggesting that behavior in this task is not influenced by a single learning system.

3.1 Monkey prefrontal cortex activity during behavioral adaptation

3.1.1 Khamassi, Quilodran, Enel, Dominey, Procyk (2014) Cerebral Cortex

Behavioral regulation and the modulation of information coding in the lateral prefrontal and cingulate cortex

Mehdi Khamassi [1,2,3,4], René Quilodran [1,2,5], Pierre Enel [1,2], Peter F. Dominey [1,2], Emmanuel Procyk [1,2]

[1] Inserm, U846, Stem Cell and Brain Research Institute, 69500 Bron, France
[2] Université de Lyon, Lyon 1, UMR-S 846, 69003 Lyon, France
[3] Institut des Systèmes Intelligents et de Robotique, Université Pierre et Marie Curie-Paris 6, F-75252, Paris Cedex 05, France
[4] CNRS UMR 7222, F-75005, Paris Cedex 05, France
[5] Escuela de Medicina, Departamento de Pre-clínicas, Universidad de Valparaíso, Hontaneda 2653, Valparaíso, Chile

Running title
Adaptive control in prefrontal cortex

Keywords:
reinforcement-learning, decision, feedback, adaptation, cingulate, reward,

This is a preprint of the paper published in *Cerebral Cortex* (Oxford University Press) in 2014.

Corresponding author:
M.K.
Institut des Systèmes Intelligents et de Robotique (UMR7222)
CNRS - Université Pierre et Marie Curie
Pyramide, Tour 55 - Boîte courrier 173
4 place Jussieu, 75252 Paris Cedex 05, France
tel: + 33 1 44 27 28 85
fax: +33 1 44 27 51 45
email: mehdi.khamassi@isir.upmc.fr

Behavioral regulation and the modulation of information coding in the lateral prefrontal and cingulate cortex

M. Khamassi, R. Quilodran, P. Enel, P.F. Dominey, E. Procyk

To explain the high level of flexibility in primate decision-making, theoretical models often invoke reinforcement-based mechanisms, performance monitoring functions, and core neural features within frontal cortical regions. However, the underlying biological mechanisms remain unknown. In recent models, part of the regulation of behavioral control is based on meta-learning principles, e.g. driving exploratory actions by varying a meta-parameter, the inverse temperature, which regulates the contrast between competing action probabilities. Here we investigate how complementary processes between lateral prefrontal cortex (LPFC) and dorsal anterior cingulate cortex (dACC) implement decision regulation during exploratory and exploitative behaviors. Model-based analyses of unit activity recorded in these two areas in monkeys first revealed that adaptation of the decision function is reflected in a covariation between LPFC neural activity and the control level estimated from the animal's behavior. Second, dACC more prominently encoded a reflection of outcome uncertainty useful for control regulation based on task monitoring. Model-based analyses also revealed higher information integration before feedback in LPFC, and after feedback in dACC. Overall the data support a role of dACC in integrating reinforcement-based information to regulate decision functions in LPFC. Our results thus provide biological evidence on how prefrontal cortical subregions may cooperate to regulate decision-making.

INTRODUCTION

When searching for resources, animals can adapt their choices by reference to the recent history of successes and failures. This progressive process leads to improved predictions of future outcomes and to the adjustment of action values. However, to be efficient, adaptation requires dynamic modulations of behavioral control, including a balance between choices known to be rewarding (exploitation), and choices with unsure, but potentially better, outcome (exploration).

The prefrontal cortex is required for the organization of goal-directed behavior (Miller and Cohen 2001; Wilson et al. 2010) and appears to play a key role in regulating exploratory behaviors (Daw N. D. et al. 2006; Cohen J. D. et al. 2007; Frank et al. 2009). The lateral prefrontal cortex (LPFC) and the

dorsal anterior cingulate cortex (dACC, or strictly speaking the midcingulate cortex, (Amiez et al. 2013)) play central roles, but it is unclear which mechanisms underlie the decision to explore and how these prefrontal subdivisions participate.

Computational solutions often rely on the meta-learning framework, where shifting between different control levels (e.g. shifting between exploration and exploitation) is achieved by dynamically tuning meta-parameters based on measures of the agent's performance (Doya 2002; Ishii et al. 2002; Schweighofer and Doya 2003). When applied to models of prefrontal cortex's role in exploration (McClure et al. 2006; Cohen J. D. *et al.* 2007; Krichmar 2008; Khamassi et al. 2011), this principle predicts that the expression of exploration is associated with decreased choice-selectivity in the LPFC (flat action probability distribution producing stochastic decisions) while exploitation is associated with increased selectivity (peaked probability distribution resulting in a winner-take-all effect). However, such online variations during decision-making have yet to be shown experimentally. Moreover, current models often restrict the role of dACC to conflict monitoring (Botvinick et al. 2001) neglecting its involvement in action valuation (MacDonald et al. 2000; Kennerley et al. 2006; Rushworth and Behrens 2008; Seo and Lee 2008; Alexander W.H. and Brown 2010; Kaping et al. 2011). dACC activity shows correlates of adjustment of action values based on measures of performance such as reward prediction errors (Holroyd and Coles 2002; Amiez et al. 2005; Matsumoto et al. 2007; Quilodran et al. 2008), outcome history (Seo and Lee 2007), and error-likelihood (Brown and Braver 2005). Variations of activities in dACC and LPFC between exploration and exploitation suggest that both structures contribute to the regulation of exploration (Procyk et al. 2000; Procyk and Goldman-Rakic 2006; Landmann et al. 2007; Rothe et al. 2011).

The present work assessed the complementarity of dACC and LPFC in behavioral regulation. We previously developed a neurocomputational model of the dACC-LPFC system to synthesize the data reviewed above (Khamassi *et al.* 2011; Khamassi et al. 2013). One important feature of the model was to include a regulatory mechanism by which the control level is modulated as a function of changes in the monitored performance. As reviewed above such a regulatory mechanism should lead to changes in prefrontal neural selectivity. This work thus generated experimental predictions that are tested here on actual neurophysiological data.

We recorded LPFC single-unit activities and made comparative model-based analyses with these data and dACC recordings that had previously been analyzed only at the time of feedback (Quilodran *et al.* 2008). We show that information related to different model variables (reward prediction errors, action values, and outcome uncertainty) are multiplexed in different trial epochs both in dACC and LPFC, with higher integration of information before the feedback in LPFC, and after the feedback in dACC. Moreover LPFC activity displays higher mutual information with the animal's choice than dACC,

supporting its role in action selection. Importantly, as predicted by prefrontal cortical models, we observe that LPFC choice selectivity co-varies with the control level measured from behavior. Taken together with recent data (Behrens et al. 2007; Rushworth and Behrens 2008), our results suggest that the dACC-LPFC diad is implicated in the online regulation of learning mechanisms during behavioral adaptation, with dACC integrating reinforcement-based information to regulate decision functions in LPFC.

MATERIAL & METHODS

Monkey housing, surgical, electrophysiological and histological procedures were carried out according to the European Community Council Directive (1986) (Ministère de l'Agriculture et de la Forêt, Commission nationale de l'expérimentation animale) and Direction Départementale des Services Vétérinaires (Lyon, France).

Experimental set up. Two male rhesus monkeys (monkeys M and P) were included in this experiment. During recordings animals were seated in a primate chair (Crist Instrument Company Inc., USA) within arm's reach of a tangent touch-screen (Microtouch System) coupled to a TV monitor. In the front panel of the chair, an opening allowed the monkey to touch the screen with one hand. A computer recorded the position and accuracy of each touch. It also controlled the presentation via the monitor of visual stimuli (colored shapes), which served as visual targets (CORTEX software, NIMH Laboratory of Neuropsychology, Bethesda, Maryland). Eye movements were monitored using an Iscan infrared system (Iscan Inc., USA).

Problem Solving task. We employed a Problem Solving task (PS task; **Fig. 1A**) where the subject has to find by trial and error which of four targets is rewarded. A typical problem started with a *Search* period where the animal performed a series of incorrect search trials (INC) until the discovery of the correct target (first correct trial, CO1). Then a *Repetition* period was imposed where the animal could repeat the same choice during a varying number of trials (between 3 and 11 trials) to reduce anticipation of the end of problems. At the end of repetition, a Signal to Change (SC*;* a red flashing circle of 8 cm in diameter at the center of screen) indicated the beginning of a new problem, i.e. that the correct target location would change with a 90% probability.

Each trial was organized as follows: a central target (lever) is presented which is referred to as trial start (ST); the animal then touches the lever to trigger the onset of a central white square which served as fixation point (FP). After an ensuing delay period of about 1.8 s (during which the monkey is required to maintain fixation on the FP), four visual target items (disks of 5mm in diameter) are presented and the FP is extinguished. The monkey then has to make a saccade towards the selected target. After the monkey has fixated on the selected target for 390 ms, all the targets turn white (go

signal), indicating that the monkey can touch the chosen target. Targets turn grey at touch for 600ms and then switch off. At offset, a juice reward is delivered after a correct touch. In the case of an incorrect choice, no reward is given, and in the next trial the animal can continue his search for the correct target. A trial is aborted in case of a premature touch or a break in eye fixation.

Behavioral data. Performance in search and repetition periods was measured using the average number of trials performed until discovery of the correct target (including first correct trial) and the number of trials performed to repeat the correct response three times, respectively. Different types of trials are defined in a problem. During search the successive trials were labeled by their order of occurrence (indices: 1, 2, 3, …, until the first correct trial). Correct trials were labeled CO1, CO2, … and COn. Arm reaction times and movement times were measured on each trial. Starting and ending event codes defined each trial.

Series of problems are grouped in sessions. A session corresponds to one recording file that contain data acquired for several hours (during behavioral sessions) to several tens of minutes (during neurophysiological recordings corresponding to one site and depth).

Electrophysiological recordings. Monkeys were implanted with a head-restraining device, and a magnetic resonance imaging-guided craniotomy was performed to access the prefrontal cortex. A recording chamber was implanted with its center placed at stereotaxic anterior level A+31. Neuronal activity was recorded using epoxy-coated tungsten electrodes. Recording sites labeled dACC covered an area extending over about 6 mm (anterior to posterior), in the dorsal bank and fundus of the anterior part of the cingulate sulcus, at stereotaxic levels superior to A+30 (**Fig. 1B**). This region is at the rostral level of the mid-cingulate cortex as defined by Vogt and colleagues (Vogt et al. 2005). Recording sites in LPFC were located mostly on the posterior third of the principal sulcus.

Data analyses

All analyses were performed using Matlab (The Mathworks, Natick, MA).

Theoretical model for model-based analysis. We compared the ability of several different computational models to fit trial-by-trial choices made by the animals. The aim was to select the best model to analyze neural data. The models tested (see list below) were designed to evaluate which among several computational mechanisms were crucial to reproduce monkey behavior in this task. The mechanisms are:

a) **Elimination of non-rewarded targets tested by the animal during the search period.** This mechanism could be modeled in many different ways, e.g. using Bayesian models or reinforcement learning models. In order to keep our results comparable and includable within the framework used by previous similar studies (e.g. Matsumoto et al., 2007; Seo and Lee, 2009; Kennerley and Walton, 2011), we used reinforcement learning models (which would work with

high learning rates – i.e. close to 1 – in this task) while noting that this would be equivalent to models performing logical elimination of non-rewarded targets or models using a Bayesian framework for elimination. This mechanism is included in Models 1-10 in the list below.

b) Progressive forgetting that a target has already been tested. This mechanism is included in Models 2-7 and 9-10.

c) Reset after the Signal to Change. This would represent information about the task structure and is included in Models 3-12. Among these models, some (i.e. Models 4,6-10) also tend not to choose the previously rewarded target (called 'shift' mechanism), and some (i.e. Models 5-10) also include spatial biases for the first target choice within a problem (called 'bias' mechanism).

d) Change in the level of control from search to repetition period (after the first correct trial). This would represent other information about the task structure and is included in Models 9 and 10 (i.e. GQLSB2β and SBnoA2β).

List of tested models:

1. Model QL (Q-learning)

We first tested a classical Q-learning (*QL*) algorithm which implements action valuation based on standard reinforcement learning mechanisms (Sutton and Barto 1998). The task involving 4 possible targets on the touch screen (upper-left: 1, upper-right: 2, lower-right: 3, lower-left: 4, **Fig. 1C**), the model had 4 possible action values (i.e. Q_1, Q_2, Q_3 and Q_4 corresponding to the respective values associated with choosing target 1, 2, 3 and 4 respectively).

At each trial, the probability of choosing target *a* was computed by a Boltzmann softmax rule for action selection:

$$P_a(t) = \frac{\exp(\beta Q_a(t))}{\sum_b \exp(\beta Q_b(t))} \quad (1)$$

where the inverse temperature meta-parameter β ($0 < \beta$) regulates the exploration level. A small β leads to very similar probabilities for all targets (flat probability distribution) and thus to an exploratory behavior. A large β increases the contrast between the highest value and the others (peaked probability distribution), and thus produces an exploitative behavior.

At the end of the trial, after choosing target a_i, the corresponding value is compared with the presence/absence of reward so as to compute a Reward Prediction Error (RPE) (Schultz et al. 1997):

$$\delta(t+1) = r(t+1) - Q_a(t) \quad (2)$$

where *r(t)* is the reward function modeled as being equal to 1 at the end of the trial in the case of success, and -1 in the case of failure. The reward prediction error signal *δ(t)* is then used to update the value associated to the chosen target:

$$Q_a(t+1) = Q_a(t) + \alpha\delta(t+1) \quad (3)$$

where α is the learning rate. Thus the *QL* model employs 2 free meta-parameters: α and β.

2. Model GQL (Generalized Q-learning)

We also tested a generalized version of Q-learning (*GQL*) (Barraclough et al. 2004; Ito and Doya 2009) which includes a forgetting mechanism by also updating values associated to each non chosen target *b* according to the following equation:

$$Q_b(t+1) = Q_b(t) + (1-\kappa)(Q_0 - Q_b(t)) \quad (4)$$

where *κ* is a third meta-parameter called the forgetting rate $(0<\kappa<1)$, and Q_0 is the initial Q-value.

3. Model GQLnoSnoB (GQL with reset of Q values at each new problem; no shift, no bias)

Since animals are over-trained on the PS task, they tend to learn the task structure: the presentation of the Signal to Change (SC) on the screen is sufficient to let them anticipate that a new problem will start and that most probably the correct target will change. In contrast, the two above-mentioned reinforcement learning models tend to repeat previously rewarded choices. We thus tested an extension of these models where the values associated to each target are reset to [0 0 0 0] at the beginning of each new problem (Model *GQLnoSnoB*).

4. Model GQLSnoB (GQL with reset including shift in previously rewarded target; no bias)

We also tested a version of the latter model where, in addition, the value associated to the previously rewarded target has a probability P_S of being reset to 0 at the beginning of the problem, P_S being the animal's average probability of shifting from the previously rewarded target as measured from the previous session $(0.85<P_S<0.95)$(**Fig. 2A- middle**). This model including the shifting mechanism is called *GQLSnoB* and has 3 free meta-parameters.

5. Model GQLBnoS (GQL with reset based on spatial biases; no shift)

In the fifth tested model (Model *GQLBnoS*), instead of using such a shifting mechanism, target Q-values are reset to values determined by the animal's spatial biases measured during search periods of the previous session; for instance, if during the previous session, the animal started 50% of search periods by choosing target *1*, 25% by choosing *2*, 15% by choosing target *3* and the rest of the time by choosing target *4*, target values were reset to $[\theta_1 ; \theta_2 ; \theta_3 ; (1-\theta_1-\theta_2-\theta_3)]$ where θ_1=0.5, θ_2=0.25 and θ_3=0.15 at each new search of the next session. In this manner, Q-values are reset using a rough estimate of choice variance during the previous session. These 3 spatial bias parameters are not considered as free meta-parameters since they were always determined based on the previous behavioral session because they were found to be stable across sessions for each monkey (**Fig. 2A-right**).

6. Model GQLSB *(GQL with reset including shift in previously rewarded target and spatial biases)*

We also tested a model which combines both shifting mechanism and spatial biases (Model GQLSB) and thus has 3 free meta-parameters.

7. Model SBnoA (Shift and Bias but the learning rate α is fixed to 1)

Since the reward schedule is deterministic (i.e. choice of the correct target provides reward with probability 1), a single correct trial is sufficient for the monkey to memorize which target is rewarded in a given problem. We thus tested a version of the previous model where elimination of non-rewarded target is done with a learning rate α fixed to 1 – i.e. no degree of freedom in the learning rate in contrast with Model GQLSB. This meta-parameter is usually set to a low value (i.e. close to 0) in the Reinforcement Learning framework to enable progressive learning of reward contingencies (Sutton and Barto 1998). With α set to 1, the model SBnoA systematically performs sharp changes of Q-values after each outcome, a process which could be closer to working memory mechanisms in the prefrontal cortex (Collins and Frank 2012). All other meta-parameters are similar as in GQLSB, including the forgetting mechanism (**Equation 4**) which is considered to be not specific to Reinforcement Learning but also valid for Working Memory (Collins and Frank, 2012). Model SBnoA has 2 free meta-parameters.

8. Model SBnoF (Shift and Bias but no α and no Forgetting)

To verify that the forgetting mechanism was necessary, we tested a model where both α and κ are set to 1. This model has thus only 1 meta-parameter: β.

9. Model GQLSB2β (with distinct exploration meta-parameters during search and repetition trials: resp. β_S *and* β_R*)*

To test the hypothesis that monkey behavior in the PS Task can be best explained by two distinct control levels during search and repetition periods, instead of using a single meta-parameter β for all trials, we used two distinct meta-parameters β_S and β_R so that the model used β_S in **Equation 1** during search trials and β_R in **Equation 1** during repetition trials. We tested these distinct search and repetition β_S and β_R meta-parameters in Model GQLSB2β which thus has 4 free meta-parameters compared to 3 in Model GQLSB.

10. Model SBnoA2β (with distinct exploration meta-parameters during search and repetition trials: resp. β_S *and* β_R*)*

Similarly to the previous model, we tested a version of Model SBnoA which includes two distinct β_S and β_R meta-parameters for search and repetition periods. Model SBnoA2β thus has 3 free meta-parameters.

11. and 12. Control models: ClockS (Clockwise search + repetition of correct target); RandS (Random search + repetition of correct target)

We finally tested 2 control models to test the contribution of the value updating mechanisms used in the previous models for the elimination of non-rewarded target (i.e. **Equation 3** with α used as a free meta-parameter in model GQLSB or set to 1 in Model SBnoA). Model *ClockS* replaces such mechanism by performing systematic clockwise searches, starting from the animal's favorite target – as measured in the spatial bias –, instead of choosing targets based on their values, and repeats the choice of the rewarded target once it finds it. Model RandS performs random searches and repeats choices of the rewarded target once it finds it.

Theoretical model optimization. To compare the ability of models in fitting monkeys' behavior during the task, (1) we first separated the behavioral data into 2 datasets so as to optimize the models on the Optimization dataset (Opt) and then perform an out-of-sample test of these models on the Test dataset (Test), (2) for each model, we then estimated the meta-parameter set which maximized the log-likelihood of monkeys' trial-by-trial choices in the Optimization dataset given the model, (3) we finally compared the scores obtained by the models with different criteria: maximum log-likelihood (LL) and percentage of monkeys' choice predicted (%) on Opt and Test datasets, BIC, AIC, Log of posterior probability of models given the data and given priors over meta-parameters (LPP).

1. Separation of optimization (Opt) and test (Test) datasets

We used a cross-validation method by optimizing models' meta-parameters on 4 behavioral sessions (2 per monkey concatenated into a single block of trials per monkey in order to optimize a single meta-parameter set per animal; 4031 trials) of the PS task, and then out of sample testing these models with the same meta-parameters on 49 other sessions (57336 trials). The out of sample test was performed to test models' generalization ability and to validate which model is best without complexity issues.

2. Meta-parameter estimation

The aim here was to find for each model M the set of meta-parameters θ which maximized the log-likelihood LL of the sequence of monkey choices in the Optimization dataset D given M and θ:

$$\theta_{opt} = \underset{\theta}{\arg\max}\{Log(P(D|M,\theta))\} \quad (5)$$

$$LL_{opt} = \underset{\theta}{\max}\{Log(P(D|M,\theta))\} \quad (6)$$

We searched for each model's LL_{opt} and θ_{opt} on the Optimization dataset with two different methods:

We first sampled a million different meta-parameters sets (drawn from prior distributions over meta-parameters such that α,κ are in [0;1], β,β_S,β_R are in -10log([0;1])). We stored the LL_{opt} score obtained for each model and the corresponding meta-parameter set θ_{opt}.

We then performed another meta-parameter search through a gradient-descent method using the *fminsearch* function in Matlab launched at multiple starting points: we started the function from all possible combinations of meta-parameters in α,κ in {0.1;0.5;0.9}, β,β_S,β_R in {1;5;35}. If this method gave a better LL score for a given model, we stored it as well as the corresponding meta-parameter set. Otherwise, we kept the best LL score and the corresponding meta-parameter set obtained with the sampling method for this model.

3. Model comparison

In order to compare the ability of the different models to accurately fit monkeys' behavior in the task, we used different criteria. As typically done in the literature, we first used the maximized log-likelihood obtained for each model on the Optimization dataset (LL_{opt}) to compute the Bayesian Information Criterion (BIC_{opt}) and Akaike Information Criterion (AIC_{opt}). We also looked at the percentage of trials of the Optimization dataset where each model accurately predicts monkeys' choice ($\%_{opt}$). We performed likelihood ratio tests to compare nested models (*e.g.* Model SBnoF and Model SBnoA).

To test models' generalization ability and to validate which model is best without complexity issues, we additionally compared models' log-likelihood on the Test dataset given the meta-parameters estimated on the Optimization dataset (LL_{test}), as well as models' percentage of trials of the Test dataset where the model accurately predicts monkeys' choice given the meta-parameters estimated on the Optimization dataset ($\%_{test}$).

Finally, because comparing the maximal likelihood each model assigns to data can result in overfitting, we also computed an estimation of the log of the posterior probability over models on the Optimization dataset (LPP_{opt}) estimated with the meta-parameter sampling method previously performed (Daw N.D. 2011). To do so, we hypothesized a uniform prior distribution over models P(M); we also considered a prior distribution for the meta-parameters given the models P(θ|M), which was the distributions from which the meta-parameters were drawn during sampling. With this choice of priors and meta-parameter sampling, LPP_{opt} can be written as:

$$LPP_{opt} = Log\left(P\left(M\middle|D\right)\right) \propto Log\left(\int_{\theta} P\left(D\middle|M,\theta\right)d\theta\right) \approx Log\left(\frac{1}{N}\sum_{i=1}^{N} P\left(D\middle|M,\theta_i\right)\right) \quad (7)$$

where N is the number of samples drawn for each model. To avoid numerical issues in Matlab when computing the exponential of large numbers, LPP_{opt} was computed in practice as:

$$LPP_{opt} = Log\left(\sum_{\theta} \exp\left(\log\left(P\left(D\middle|M,\theta\right)\right) - LL_{opt}\right)\right) - Log(N) + LL_{opt} \quad (8)$$

Estimating models' posterior probability given the data can be seen as equivalent as computing a "mean likelihood". And it has the advantage of penalizing both models that have a peaked posterior probability distribution (i.e. models with a likelihood which is good at its maximum but which decreases sharply as soon as meta-parameters slightly change) and models that have a large number of free meta-parameters (Daw N.D. 2011).

Neural data analyses

Activity variation between search and repetition. To analyze activity variations of individual neurons between the search period and the repetition period, we computed an index of activity variation for each cell:

$$I_a = \frac{(B - A)}{(A + B)} \quad (9)$$

A is the cell mean firing rate during the early-delay epoch ([start+0.1s; start+1.1s]) over all trials of the search period, and B is the cell's mean firing rate in the same epoch during all trials of the repetition period.

To measure significant increases or decreases of activity in a given group of neurons, we considered the distribution of neurons' activity variation index. An activity variation was considered significant when the distribution had a mean significantly different from 0 using a one-sample t-test and a median significantly different from zero using a Wilcoxon Mann-Whitney U-test for zero median. Then we employed a Kruskal-Wallis test to compare the distributions of activity during search and repetition, corrected for multiple comparison between different groups of neurons (Bonferroni correction).

Choice selectivity. To empirically measure variations in choice selectivity of individual neurons, we analyzed neural activities using a specific measure of spatial selectivity (Procyk and Goldman-Rakic 2006). The activity of a neuron was classified as choice selective when this activity was significantly modulated by the identity/location of the target chosen by the animal (one-way ANOVA, $p < 0.05$). The target preference of a neuron was determined by ranking the average activity measured in the early-delay epoch ([start+0.1s; start+1.1s]) when this activity was significantly modulated by the target choice. We used for each unit the average firing rate ranked by values and herein named 'preference' (a, b, c, d where a is the preferred and d the least preferred target). The ranking was first

used for population data and structure comparisons. For each cell, the activity was normalized to the maximum and minimum of activity measured in the repetition period (with normalized activity = [activity - min]/[max - min]).

Second, to study changes in choice selectivity (tuning) throughout trials during the task, we used for each unit the average firing rate ranked by values (a, b, c, d). We then calculated the norm of a preference vector using the method of (Procyk and Goldman-Rakic 2006) which is equivalent to computing the Euclidean distance within a factor of $\sqrt{2}$: We used an arbitrary arrangement in a square matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ to calculate the vector norm:

$$H=(a+c)-(b+d) \text{ and } V=(a+b)-(c+d) \qquad (10)$$

$$\text{norm} = \sqrt{H^2 + V^2}$$

For each neuron, the norm was divided by the global mean activity of the neuron (to exclude the effect of firing rate in this measure: preventing a cell A that has a higher mean firing rate than a cell B to have a higher choice selectivity norm when they are both equally choice selective).

The value of the preference vector norm was taken as reflecting the strength of choice coding of the cell. A norm equal to zero would reflect equal activity for the four target locations. This objective measure allows the extraction of one single value for each cell, and can be averaged across cells. Finally, to study variations in choice selectivity between search and repetition periods, we computed an index of choice selectivity variation for each cell:

$$I_s = \frac{(D-C)}{(C+D)} \qquad (11)$$

where C is the cell's choice selectivity norm during search and D is the cell's choice selectivity norm during repetition.

To assess significant variations of choice selectivity between search and repetition in a given group of neurons (*e.g.* dACC or LPFC), we used: a t-test to verify whether the mean was different from zero; a Wilcoxon Mann-Whitney U- test to verify whether the median was different from zero; then we used a Kruskal-Wallis test to compare the distributions of choice selectivity during search and repetition, corrected for multiple comparison between different groups of neurons (Bonferroni correction).

To assess whether variations of choice selectivity between search and repetition depended on the exploration level β measured in the animal's behavior by means of the model, we cut sessions into two groups: those where β was smaller than the median of β values (i.e. 5), and those where β was larger than this median. Thus, in these analyses, repetition periods of a session with $\beta < 5$ will be considered a relative exploration, and repetition periods of a session with $\beta > 5$ will be considered a relative exploitation. We then performed two-way ANOVAs (β x task phase) and used a Tukey HSD

post hoc test to determine the direction of the significant changes in selectivity with changing exploration levels, tested at p=0.05.

Model-based analysis of single-unit data. To test whether single units encoded information related to model computations, we used the following model variables as regressors of trial-by-trial activity: the reward prediction error [δ], the action value [Q] associated to each target and the outcome uncertainty [U]. The latter is a performance monitoring measure which assesses the entropy of the probability over the different possible outcomes (i.e. reward r versus no reward $\bar{r}$) at the current trial *t* given the set T of remaining targets: $U(t) = -P(r|T)\log(P(r|T)) - P(\bar{r}|T)\log(P(\bar{r}|T))$. At the beginning of a new problem, when there are 4 possible targets, U starts at a low value since there is 75% chance of making an error. U increases trial after trial during the search period. It is maximal when there remain 2 possible targets because there is 50% chance of making an error. Then U drops after either the first rewarded trial or the third error trial – because the fourth target is necessarily the rewarded one – and remains at zero during the repetition period. We decided to use a regressor with this pattern of change because it is somewhat comparable to the description of changes in frontal activity previously observed during the PS task (Procyk et al., 2000; Procyk and Goldman-Rakic, 2006).

We used U as the simplest possible parameter-free performance monitoring regressor for neural activity. This was done in order to test whether dACC and LPFC single-unit could reflect performance monitoring processes in addition to responding to feedback and tracking target values. But we note that the profile of U in this task would not be different from other performance monitoring measures such as the outcome history that we previously used in our computational model for dynamic control regulation in this task (Khamassi *et al.* 2011), or such as the vigilance level in the model of Dehaene and Changeux (Dehaene et al. 1998) which uses error and correct signals to update a regulatory variable (increased after errors and decreased after correct trials). We come back to possible interpretations of neural correlates of U in the discussion.

To investigate how neural activity was influenced by action values [Q], reward prediction errors [δ] as well as the outcome uncertainty [U], we performed a multiple regression analysis combined with a bootstrapping procedure, focusing our analyses on spike rates during a set of trial epochs (**Fig. 1C**): pre-start (0.5 s before trial start); post-start (0.5 s after trial start); pre-target (0.5 s before target onset); post-target (0.5 s after target onset); the action epoch defined as pre-touch (0.5 s before screen touch); pre-feedback (0.5 s before feedback onset); early-feedback (0.5 s after feedback onset); late-feedback (1.0 s after feedback period); inter-trial-interval (ITI; 1.5 s after feedback onset).

The spike rate *y(t)* during each of these intervals in trial *t* was analyzed using the following multiple linear regression model:

$$y(t) = \rho_0 + \rho_1 Q_1(t) + \rho_2 Q_2(t) + \rho_3 Q_3(t) + \rho_4 Q_4(t) + \rho_5 \delta(t) + \rho_6 U(t) \quad (13)$$

where $Q_k(t), (k \in \{1..4\})$ are the action values associated to the four possible targets at time t, δ(t) is the reward prediction error, U(t) is the outcome uncertainty, and $\rho_i, (i \in \{1..n\})$ are the regression coefficients.

δ, Q and U were all updated once in each trial. δ was updated at the time of feedback, so that regression analyses during pre-feedback epochs were done using δ from the previous trial, while analyses during post-feedback epochs used the updated δ. Q and U were updated at the end of the trial so that regression analyses in all trial epochs were done using the Q-values and U value of the current trial.

Note that the action value functions of successive trials are correlated, because they are updated iteratively, and this violates the independence assumption in the regression model. Therefore, the statistical significance for the regression coefficients in this model was determined by a permutation test. For this, we performed a shuffled permutation of the trials and recalculated the regression coefficients for the same regression model, using the same meta-parameters of the model obtained for the unshuffled trials. This shuffling procedure was repeated 1000 times (bootstrapping method), and the *p* value for a given independent variable was determined by the fraction of the shuffles in which the magnitude of the regression coefficient from the shuffled trials exceeded that of the original regression coefficient (Seo and Lee 2009), corrected for multiple comparisons with different model variables in different trial epochs (Bonferroni correction).

To assess the quality of encoding of action value information by dACC and LPFC neurons, we also performed a multiple regression analysis on the activity of each neuron related to Q-values after excluding trials where the preferred target of the neuron was chosen by the monkey. This analysis was performed to test whether the activity of such neurons still encodes Q-values outside trials where the target is selected. Similarly, to evaluate the quality of reward prediction error encoding, we performed separate multiple regression analyses on correct trials only versus error trials only. This analysis was performed to test whether the activity of such neurons quantitatively discriminate between different amplitudes of positive reward prediction errors and between different amplitudes of negative reward prediction errors. In both cases, the significance level of the multiple regression analyses was determined with a bootstrap method and a Bonferroni correction for multiple comparisons.

Finally, to measure possible collinearity issues between model variables used as regressors of neural activity, we used Brian Lau's Collinearity Diagnostics Toolbox for Matlab (http://www.subcortex.net/research/code/collinearity-diagnostics-matlab-code (Lau 2014)). We extracted the variation inflation factors (VIF) computed with the coefficient of determination obtained

when each regressor was expressed as a function of the other regressors. We also computed the condition indexes (CONDIND) and variance decomposition factors (VDF) obtained in the same analysis. A strong collinearity between regressors was diagnosed when CONDIND ≥ 30 and more than two VDFs > 0.5. A moderate collinearity was diagnosed when CONDIND ≥ 10 and more than two VDFs > 0.5. CONDIND ≤ 10 indicated a weak collinearity.

Principal component analysis. To determine the degree to which single-unit activity segregated or integrated information about model variables, we performed a Principal Component Analysis (PCA) on the 3 correlation coefficients $\rho_i, (i \in \{4..6\})$ obtained with the multiple regression analysis and relating neural activity with the 3 main model variables (reward prediction error δ, outcome uncertainty U, and the action value Q_k associated to the animal's preferred target k). For each trial epoch, we pooled the coefficients obtained for all neurons in correlation with these model variables. Each principal component being expressed as a linear combination of the vector of correlation coefficients of neuron activities with these three model variables, the contribution of different model variables to each component gives an idea as to which extent cell activity is explained by an integrated contribution of multiple model variables. For instance, if a PCA on cell activity in the early-delay period produces three principal components that are each dependent on a different single model variable (*e.g.* PC1 = 0.95Q + 0.01δ + 0.04U; PC2 = 0.1Q + 0.8δ + 0.1U; PC3 = 0.05Q + 0.05δ + 0.9U), then activity variations are best explained by separate influences from the information conveyed by the model variables. If in contrast, the PCA produces principal components which strongly depend on multiple variables (*e.g.* PC1 = 0.5Q + 0.49δ + 0.01U; PC2 = 0.4Q + 0.1δ + 0.5U; PC3 = 0.2Q + 0.4δ + 0.4U), then variations of the activities are best explained by an integrated influence of such information (see **Supplementary Figure S1** for illustration of different Principal Components resulting from artificially generated data showing different levels of integration between model variables).

We compared the normalized absolute values of the coefficients of the three principal components so that a coefficient close to 1 denotes a strong correlation while a coefficient close to 0 denotes no correlation. To quantify the integration of information about different model variables in single-unit activities, for each neuron *k*, we computed an entropy-like index (ELI) of sharpness of encoding of different model variables based on the distributions of regression coefficients between cell activities and model variables:

$$ELI_k = -\sum_i c_i \log(c_i) \qquad (14)$$

Where c_i is the absolute value of the z-scored correlation strength ρ_i with model variable *i*. A neuron with activity correlated with different model variables with similar strengths will have a high

ELI; a neuron with activity highly correlated with only one model variable will have a low ELI. We compared the distributions of ELIs between dACC and LPFC in each trial epoch using a Kruskal-Wallis test.

Finally, we estimated the contribution of each model variable to neural activity variance in each epoch and compared it between dACC and LPFC. To do so, we first normalized the coefficients for each principal component in each epoch. These coefficients being associated to three model variables Q, δ and U, this provided us with a contribution of each model variable to each principal component in each epoch. We then multiplied them by the contribution of each principal component to the global variance in neural activity in each epoch. The result constituted a normalized contribution of each model variable to neural activity variance in each epoch. We finally computed the entropy-like index (ELI) of these contributions. We compared the set of epoch-specific ELI between dACC and LPFC with a Kruskal-Wallis test.

Mutual information. We measured the mutual information between monkey's choice at each trial and the firing rate of each individual recorded neuron during the early-delay epoch ([ST+0.1s; ST+1.1s]). The mutual information $I(S;R)$ was estimated by first computing a confusion matrix (Quian Quiroga and Panzeri 2009), relating at each trial *t*, the spike count from the unit activity in the early-delay epoch (as "predicting response" *R*) and the target chosen by the monkey (*i.e.* 4 targets as "predicted stimulus" *S*). Since neuronal activity was recorded during a finite number of trials, not all possible response outcomes of each neuron to each stimulus (target) have been sufficiently sampled. This is called the "limited sampling bias" which can be overcome by subtracting a correction term from the plug-in estimator of the mutual information (Panzeri et al. 2007). Thus we subtracted the Panzeri Treves (PT) correction term (Treves and Panzeri 1995) from the estimated mutual information $I(S;R)$:

$$BIAS\left(I\left(S;R\right)\right)=\frac{1}{2N\ln(2)}\left(\sum_{s}\left(\overline{R}_{s}-1\right)-\left(\overline{R}-1\right)\right) \quad (15)$$

Where *N* is the number of trials during which the unit activity was recorded, $\bar{R}$ is the number of relevant bins among the *M* possible values taken by the vector of spike counts and computed by the "bayescount" routine provided by (Panzeri and Treves 1996), and $\overline{R_S}$ is the number of relevant responses to stimulus (target) *s*.

Such measurement of information being reliable only if the activity was recorded during a sufficient number of trials per stimulus presentation, we restricted this analysis to units that verified the following condition (Panzeri *et al.* 2007):

$$N_S / \bar{R} \geq 4 \quad (16)$$

Where N_S is the minimum number of trials per stimulus (target).

Finally, to verify that such a condition was sufficiently restrictive to exclude artifactual effects, for each considered neuron we constructed 1000 pseudo response arrays by shuffling the order of trials at fixed target stimulus, and we recomputed each time the mutual information in the same manner (Panzeri *et al.* 2007). Then we verified that the average mutual information obtained with such shuffling procedure was close to the PT bias correction term computed with **Equation 15** (Panzeri and Treves 1996).

RESULTS

Previous studies have emphasized the role of LPFC in cognitive control and dACC in adjustment of action values based on measures of performance such as reward prediction errors, error-likelihood and outcome history. In addition, variations of activities in the two regions between exploration and exploitation suggest that both contribute to the regulation of the control level during exploration. Altogether neurophysiological data suggest particular relationships between dACC and LPFC, but their respective contribution during adaptation remains unclear and a computational approach to this issue appears highly relevant. We recently modeled such relationships using the meta-learning framework (Khamassi *et al.* 2011). The network model was simulated in the Problem Solving (PS) task (Quilodran et al., 2008) where monkeys have to search for the rewarded target in a set of four on a touch-screen, and have to repeat this rewarded choice for at least 3 trials before starting a new search period (**Fig. 1A**). In these simulations, variations of the model's control meta-parameter (i.e. inverse temperature β) produced variations of choice selectivity in simulated LPFC in the following manner: a decrease of choice selectivity (exploration) during search; an increase of choice selectivity (exploitation) during repetition. This resulted in a globally higher mean choice selectivity in simulated LPFC compared to simulated dACC, and in a co-variation between choice selectivity and the inverse temperature in simulated LPFC but not in simulated dACC (Khamassi *et al.* 2011). This illustrates a prediction of computational models on the role of prefrontal cortex in exploration (McClure *et al.* 2006; Cohen J. D. *et al.* 2007; Krichmar 2008) which has not yet been tested experimentally.

Characteristics of behaviors

To assess the plausibility of such computational principles we first analyzed animals' behavior in the PS task. During recordings, monkeys performed nearly optimal searches, *i.e.*, rarely repeated incorrect trials (INC), and on average made errors in less than 5% of repetition trials. Although the animals' strategy for determining the correct target during search periods was highly efficient, the pattern of successive choices was not systematic. Analyses of series of choices during search periods revealed that monkeys used either clockwise (e.g. choosing target 1 then 2), counterclockwise, or

crossing (going from one target to the opposite target in the display, e.g. from 1 to 3) strategies, with a slightly higher incidence for clockwise and counterclockwise strategies, and a slightly higher incidence for clockwise over counterclockwise strategy (Percent clockwise, counterclockwise, crossing and repeats were 38%, 36%, 25%, 1% and 39%, 33%, 26%, 2% for each monkey respectively, measured for 9716 and 4986 transitions between two targets during search periods of 6986 and 3227 problems respectively). Rather than being systematic or random, monkeys' search behavior appeared to be governed by more complex factors: shifting from the previously rewarded target in response to the Signal to Change (SC) at the beginning of most new problems (**Fig. 2A-middle**); spatial biases *i.e.* more frequent selection of preferred targets in the first trial of search periods (**Fig. 2A-right**); and efficient adaption to each choice error as argued above. This indicates a planned and controlled exploratory behavior during search periods. This is also reflected in an incremental change in reaction times during the search period, with gradual decreases after each error (**Fig. 2B)**. Moreover, reaction times shifted from search to repetition period after the first reward (CO1), suggesting a shift between two distinct behavioral modes or two levels of control (Monkey M: Wilcoxon Mann-Whitney U-test, $p < 0.001$; Monkey P: $p < 0.001$; **Fig. 2B**).

Model-based analyses. Behavioral analyses revealed that monkeys used nearly-optimal strategies to solve the task, including shift at problem changes, which are unlikely to be solved by simple reinforcement learning. In order to identify the different elements that took part in monkey's decisions and adaptation during the task we compared the fit scores of several distinct models to trial-by-trial choices after estimating each model's free meta-parameters that maximize the log-likelihood separately for each monkey (see Methods). We found that models performing either a random search or a clockwise search and then simply repeating the correct target could not properly reproduce monkeys' behavior during the task, even when the clockwise search was systematically started by the monkeys' preferred target according to its spatial biases (Models RandS and ClockS; **Table 1** and **Fig. 2D**). Moreover, the fact that monkeys most often shifted their choice at the beginning of each new problem in response to the Signal to Change (SC) (**Fig. 2A-middle**) prevented a simple reinforcement learning model (Q-learning) or even a generalized reinforcement learning model from reproducing monkey's behavior (resp. QL and GQL in **Table 1**). Indeed, these models obviously have a strong tendency to choose the previously rewarded target without taking into account the Signal to Change to a new problem. Behavior was better reproduced with a combination of generalized reinforcement learning and reset of target values at each new problem (shifting the previously rewarded target and taking into account the animal's spatial biases measured during the previous session; i.e. Models GQLSB, GQLSB2β, SBnoA, SBnoA2β in **Figure 2D** and **Table 1**). We tested

control models without spatial biases, without problem shift, and with neither of them, to show that they were both required to fit behavior (resp. GQLSnoB, GQLBnoS and GQLnoSnoB in **Table 1**). We also tested a model with spatial biases and shift but without progressive updating of target values nor forgetting – i.e. $\alpha = 1, \kappa = 1$ (Model SBnoF, which is a restricted and nested version of Model SBnoA with 1 less meta-parameter) and found that it was not as good as SBnoA in fitting monkeys' behavior, as found with a likelihood ratio test at p=0.05 with one degree of freedom.

Although Models GQLSB, GQLSB2β, SBnoA, SBnoA2β were significantly better than other tested models along all used criteria (maximum likelihood [Opt-LL], BIC score, AIC score, log of posterior probability [LPP], out-of-sample test [Test-LL] in **Table 1**), these 4 versions gave similar fit performance. In addition, the best model was not the same depending on the considered criterion: Model GQLSB2β was the best according to LL, BIC and AIC scores, and second best according to LPP and Test-LL scores; Model SBnoA2β was the best according to LPP score; Model GQLSB was the best according to Test-LL score.

As a consequence, the present dataset does not allow to decide whether allowing a free meta-parameter α (i.e. learning rate) in model GQLSB and GQLSB2β is necessary or not in this task, compared to versions of these models where α is fixed to 1 (Model SBnoA and SBnoA2β) (**Fig. 2D** and **Table 1**). This is due to the structure of the task – where a single correct trial is sufficient to know which is the correct target – which may be solved by sharp updates of working memory rather than by progressive reinforcement learning (although a small subset of the sessions were better fitted with $\alpha \in [0.3;0.9]$ in Model GQLSB, thus revealing a continuum in the range of possible αs, **Supplementary Fig. S2**). We come back to this issue in the discussion.

Similarly, models that use distinct control levels during search and repetition (Models GQLSB2β and SBnoA2β) could not be distinguished from models using a single parameter (Models GQLSB and SBnoA) in particular because of out-of-sample test scores (**Table 1**).

Nevertheless, model-based analyses of behavior in the PS task suggest complex adaptations possibly combining rapid updating mechanisms (i.e. α close to 1), forgetting mechanisms and the use of information about the task structure (Signal to Change; first correct feedback signaling the beginning of repetition periods). Model GQLSB2β here combines these different mechanisms in the more complete manner and moreover won the competition against the other models according to three criteria out of five. Consequently, in the following we will use Model GQLSB2β for model-based analyses of neurophysiological data and will systematically compare the results with analyses performed with Models GQLSB, SBnoA, SBnoA2β to verify that they yield similar results.

In summary, the best fit was obtained with Models SBnoA, SBnoA2β, GQLSB, GQLSB2β which could predict over 80% of the choices made by the animal (**Table 1**). **Figure 2A** shows a sample of

trials where Model SBnoA can reproduce most monkey choices, and illustrating the sharper update of action values in Model SBnoA (with $\alpha = 1$) compared to Model GQLSB (where the optimized $\alpha = 0.7$). When freely simulated on 1000 problems of the PS task – i.e., the models learned from their own decisions rather than trying to fit monkeys' decisions –, the models made 38.23% clockwise search trials, 32.41% counter-clockwise, 29.22% crossing and 0.15% repeat. Simulations of the same models without spatial biases produced less difference between percentages of clockwise, counter-clockwise and crossing trials, unlike monkeys: 33.98% clockwise, 32.42% counter-clockwise, 33.53% crossing and 0.07% repeat.

Distinct control levels between search and repetition. To test whether behavioral adaptation could be described by a dynamical regulation of the β meta-parameter (*i.e.* inverse temperature) between search and repetition, we analyzed the value of the optimized two distinct free meta-parameters (β_S and β_R) in Models GQLSB2β and SBnoA2β (**Fig. 2E, 2C** and **Suppl. Fig. S2**). The value of the optimized β_S and β_R meta-parameters obtained for a given monkey in a given session constituted a quantitative measure of the control level during that session. Such level was non-linearly linked to the number of errors the animal made. For instance, a β_R of 3, 5, or 10 corresponded to approximately 20%, 5%, and 0% errors respectively made by the animal during repetition periods (**Fig. 2C**).

Interestingly, the distributions of β_S and β_R obtained for each recording session showed dissociations between search and repetition periods in a large number of sessions. We found a unimodal distribution for the β meta-parameter during the search period (β_S), reflecting a consistent level of control in the animal behavior from session to session. In contrast, we observed a bimodal distribution for the β meta-parameter during the repetition period (β_R; **Fig. 2E**). In **Figure 2E,** the peak on the right of the distribution (large β_R) corresponds to a subgroup of sessions where behavior shifted between different control levels from search to repetition periods. This shift in the level of control could be interpreted as a shift from exploratory to exploitative behavior, an attentional shift or a change in the working memory load, as we discuss further in the Discussion. Nevertheless this is consistent with the hypothesis of a dynamical regulation of the inverse temperature β between search and repetition periods in this task (Khamassi *et al.* 2011; Khamassi *et al.* 2013). The bimodal distribution for β_R illustrates the fact that during another subgroup of sessions (small β_R), the animal's behavior did not shift to a different control level during repetition and thus made more errors. Such bimodal distribution of the β meta-parameter enables to separate sessions in two groups and to compare dACC and LPFC activities (see below) during sessions where decisions displayed a shift and during sessions where no such clear shift occurred. Interestingly, the bimodal distribution of β_R is not

crucially dependent of the optimized learning rate α since a similar bimodal distribution was obtained with Model SBnoA2β and since the optimized β_S and β_R values in the two models were highly correlated (N = 277; β_S: r = 0.9, p < 0.001; β_R: r = 0.96, p < 0.001; **Supplementary Fig. S2**).

Modulation of information coding

To evaluate whether a behavioral change between search and repetition was accompanied by changes in LPFC activity and choice selectivity, we analyzed a pool of 232 LPFC single-units (see **Fig. 1B** for the anatomy) in animals performing the PS task, and compared the results with 579 dACC single-unit recordings which have been only partially used for investigating feedback-related activity (Quilodran *et al.* 2008). We report here a new study relying on comparative analyses of dACC and LPFC responses, the analysis of activities before the feedback – especially during the delay period –, and the model-based analysis of these neurophysiological data. The results are summarized in **Supplementary Table 1**.

Average activity variations between search and repetition. Previous studies revealed differential prefrontal fMRI activations between exploitation (where subjects chose the option with maximal value) and exploration trials (where subjects chose a non-optimal option) (Daw N. D. *et al.* 2006). Here a global decrease in average activity level was also observed in the monkey LPFC from search to repetition. For early-delay activity, the average index of variation between search and repetition in LPFC was negative (mean: -0.05) and significantly different from zero (mean: t-test $p < 0.001$, median: Wilcoxon Mann-Whitney U- test $p < 0.001$). The average index of activity variation in dACC was not different from zero (mean: -0.008; t-test $p > 0.35$; median: Wilcoxon Mann-Whitney U- test $p > 0.25$). However, close observation revealed that the non-significant average activity variation in dACC was due to the existence of equivalent proportions of dACC cells showing activity increase or activity decrease from search to repetition, leading to a null average index of variation (**Fig. 3A-B**; 17% versus 20% cells respectively). In contrast, more LPFC single units showed a decreased activity from search to repetition (18%) than an increase (8%), thus explaining the apparent global decrease of average LPFC activity during repetition. The difference in proportion between dACC and LPFC is significant (Pearson χ^2 test, 2 df, t = 13.0, $p < 0.01$) and was also found when separating data for the two monkeys (**Supplementary Fig. S3**). These changes in neural populations thus suggest that global non-linear dynamical changes occur in dACC and LPFC between search and repetition instead of a simple reduction or complete cessation of involvement during repetition.

Modulations of choice selectivity between search and repetition. As shown in **Figure 3A**, a higher proportion of neurons showed a significant choice selectivity in LPFC (155/230, 67%) than in dACC

(286/575, 50%; Pearson χ^2 test, 1 df, t = 20.7, p < 0.001) – as measured by the vector norm in **Equation 10**. Interestingly, the population average choice selectivity was higher in LPFC (0.80) than in dACC (0.70; Kruskal-Wallis test, p < 0.001; see **Fig. 3C**). When pooling all sessions together, this resulted in a significant increase in average choice selectivity in LPFC from search to repetition (mean variation: 0.04; Wilcoxon Mann-Whitney U-test p < 0.01; t-test p < 0.01; **Fig. 3C**).

Strikingly, the significant increase in LPFC early-delay choice selectivity from search to repetition was found only during sessions where the model fit dissociated control levels in search and repetition (i.e. sessions with large β_R [β_R > 5]; Kruskal-Wallis test, 1df, χ^2 = 6.45, p = 0.01; posthoc test with Bonferroni correction indicated that repetition > search). Such an effect was not found during sessions where the model reproducing the behavior remained at the same control level during repetition (i.e. sessions with small β_R [β_R < 5]; Kruskal-Wallis test, p > 0.98) (**Fig. 4-bottom**).

Interestingly, choice selectivity in LPFC was significantly higher during repetition for sessions where β_R was large (mean choice selectivity = 0.91) than for sessions where β_R was small (mean choice selectivity = 0.73; Kruskal-Wallis test, 1df, χ^2 = 12.5, p < 0.001; posthoc test with Bonferroni correction; **Fig. 4-bottom**). Thus, LPFC early-delay choice selectivity clearly covaried with the level of control measured in the animal's behavior by means of the model.

There was also an increase in dACC early-delay choice selectivity between search and repetition consistent with variations of β, but only during sessions where the model capturing the animal's behavior made a strong shift in the control level (β_R > 5; mean variation = 0.035, Kruskal-Wallis test, 1df, χ^2 = 5.22, p < 0.05; posthoc test with Bonferroni correction indicated that repetition > search; **Fig. 4-top**). However, overall, dACC choice selectivity did not follow variations of the control level. Two-way ANOVAs either for (β_S x task phase) or for (β_R x task phase) revealed no main effect of β (p > 0.2), an effect of task period (p < 0.01), but no interaction (p > 0.5). And there was no significant difference in ACC choice selectivity during repetition between sessions with a large β_R (mean choice selectivity = 0.69) and sessions with a low one (mean choice selectivity = 0.75; Kruskal-Wallis test, 1 df, χ^2 = 3.11, p > 0.05).

At the population level, increases in early-delay mean choice selectivity from search to repetition were due both to an increase of single unit selectivity, and to the emergence in repetition of selective units that were not significantly so in search (**Fig. 3A**). Importantly, the proportion of LPFC early-delay choice selective neurons during repetition periods of sessions where β_R was small (55%) was significantly smaller than the proportion of such LPFC neurons during sessions where β_R was large (72%; Pearson χ^2 test, 1 df, t = 7.19, p < 0.01). In contrast, there was no difference in proportion of dACC early-delay choice selective neurons during repetition between sessions where β_R was small (38%) and sessions where β_R was large (35%; Pearson χ^2 test, 1 df, t = 0.39, p > 0.5; **Fig. 4B**). These

analyses thus show a significant difference between dACC and LPFC neural activity properties. LPFC mean choice selectivity as well as LPFC proportion of choice selective cells varied between search and repetition in accordance with the control level measured in the behavior by means of the computational model, while such effect was much weaker in dACC. These results are robust since they could also be obtained with Model SBnoA2β (**Supplementary Fig. S4A**). Data separated for the two monkeys also reflected the contrast between the two structures (**Supplementary Fig. S4B**).

Mutual information between neural activity and target choice. Generally, computational models of the dACC-LPFC system make the assumption that LPFC is central for the decision output. LPFC activity should thus be more tightly related to the animal's choice than dACC activity. Here, in 63 LPFC neurons recorded during a sufficient number of presentations of each target choice (see Methods), the average mutual information – corrected for sampling bias – was more than twice as high (I_{LPFC} = 0.10 bit) as in 85 dACC cells (I_{ACC} = 0.04 bit; Kruskal-Wallis test, $p < 0.001$) (**Fig. 3D**). This effect appeared to be the result of the activity of a small subset of LPFC activity – in both monkeys (**Supplementary Fig. S3D**) – with a high mutual information with choice. To verify that the applied restriction on the number of sampling trials was accurate, we constructed 1000 shuffled pseudo response arrays for each single unit and measured the average mutual information obtained with this shuffling procedure. For the 63 LPFC and 85 dACC selected neurons, the difference between the averaged shuffled information and the bias correction term was very small (mean=0.01 bit), while it was high in non-selected neurons (mean=0.08 bit). Thus the difference in estimated information between dACC and LPFC was not due to a limited sampling bias in the restricted number of analyzed neurons. We can conclude that, in agreement with computational models of the dACC-LPFC system, neural recordings show a stronger link between LPFC activity and choice than between dACC activity and choice.

Neural activity correlated with model variables.

Following model-based analyses of behavior we tested whether single unit activity in LPFC and dACC differentially reflect information similar to variables in Model GQLSB2β by using the time series of these variables as regressors in a general linear model of single-unit activity (multiple regression analysis with a bootstrapping control – see Methods) (**Fig. 6**). In dACC and LPFC, respectively 397/579 (68.6%) cells and 145/232 (62.5%) cells showed a correlation with at least one of the model's variables in at least one of the behavioral epochs: pre-start, delay, pre-target, post-target, pre-touch, pre-feedback, early-feedback, late-feedback, and inter-trial interval (ITI). More precisely, we found a larger proportion of cells in LPFC than in dACC correlated with at least one model variable in the

post-target epoch (**Fig. 6E**; Pearson χ^2 tests, T = 3.89, p < 0.05), and a larger proportion of cells in dACC than in LPFC correlated with at least one model variable in the early-feedback epoch (Pearson χ^2 test, T = 7.90, p < 0.01). Differences in proportions of LPFC and dACC neurons correlated with different model variables during pre- or post-feedback epochs were also observed for the two monkeys separately (**Supplementary Figure S6**), and when the model-based analysis was done with Models GQLSB, SBnoA or SBnoA2β (**Supplementary Figures S5**). Collinearity diagnostics between model variables revealed a weak collinearity in 306/308 recording sessions, a moderate collinearity in 1 session and a strong collinearity in 1 session (**Supplementary Figure S9**), thus excluding the possibility that these results could be an artifact of collinearity between model variables.

Figure 5A shows an example dACC post-target activity negatively correlated with the action value associated to choosing target #4 (**Fig. 5A-top**). The raster plot and peristimulus histogram for this activity show lower firing rate in trials where the animal chose target #4 than in trials where he chose one of the other targets (**Fig. 5A-middle**). Plotting the trial-by-trial evolution of the post-target firing rate of the neuron reveals sharp variations following action value update and distinct from the time series of the other model variables δ and U (**Fig. 5A-bottom**). The firing rate dropped below baseline during trials where target #4 was chosen. Strikingly, the firing rate sharply increased above baseline in trials following non-rewarded choices of target #4. Thus this single unit not only responded when the animal selected the associated target but also kept track of the stored value associated with that target. **Figure 5B** shows a LPFC unit whose activity in the post-target epoch is positively correlated with the action value associated to choosing target #2. The raster plot illustrates a higher firing rate for trials where target #2 was chosen (grey histogram and raster, **fig. 5B-middle**). Similarly to the previous example, the trial-by-trial evolution of the post-target firing rate reveals sharp variations from trial to trial (**Fig. 5B-bottom**), consistent with sharp changes of action values in the model that best described behavior adaptation in this task (**Fig. 2A**).

We found 126/145 (87%) LPFC and 227/397 (57%) dACC Q-value encoding cells. The proportion was significantly greater in LPFC (Pearson χ2 test, 1 df, T = 41.30, p < 0.001; **Fig. 6A**). We next verified whether the activity of these cells carried Q value information only during trials where the neuron's preferred target was selected by the monkey, or also during other trials. To do so, we performed a new multiple regression analysis on the activity of each cell after excluding trials where the cell's preferred target was chosen. The activity of respectively 18% (23/126) and 13% (29/227) of LPFC and dACC Q value encoding cells were still significantly correlated with a Q value in the same epoch after excluding trials where the cell's preferred target was selected by the animal (multiple regression analysis with Bonferroni correction). Importantly, the difference in proportion of Q cells between LPFC

and dACC was still significant after restricting to Q cells showing a significant correlation while excluding trials with their preferred target (LPFC: 23/145, 16%; dACC: 29/397, 7%; Pearson $\chi 2$ test, 1 df, T = 8.97, $p < 0.01$).

Given the deterministic nature of the task, and thus the limited sampling of options, a question remains of whether these neurons really encode Q values or whether they participate to action selection. The control analysis above excluding trials with each cells' preferred target showed that at least a certain proportion of these cells carried information about action values outside trials where the corresponding action is selected. But how much information about choice do these neurons carry and is there a quantitative difference between LPFC and dACC? Interestingly, 43% (54/126) of LPFC Q cells had high mutual information with monkey choice ($I > 0.1$) whereas only 33% (75/227) of dACC Q cells verified such condition. The difference in proportion was marginally significant (Pearson $\chi 2$ proportion test, 1df, T = 3.37, $p = 0.07$). Moreover, LPFC Q cells activity contained more information about monkey choice (mean I = 0.12) than dACC Q cells (mean I = 0.09; Kruskal-Wallis test, 1df, $\chi 2 = 3.88$, $p < 0.05$; Posthoc test with Bonferroni correction found that LPFC-Q > dACC-Q) and more than LPFC non-Q cells (average = 0.09; Kruskal-Wallis test, $\chi 2 = 6.65$, 1df, $p < 0.01$; Posthoc test with Bonferroni correction found that LPFC-Q > LPFC-nonQ). dACC Q cells activity did not contain more information about monkey choice than LPFC non-Q cells (Kruskal, 1df, $\chi 2 = 1.57$, $p > 0.05$). Although the observed difference in Q-encoding between dACC and LPFC are weak, these results are in line with the hypothesized dACC role in action value encoding and with the transfer of such information to LPFC for action selection – the LPFC would encode a probability distribution over possible actions.

Feedback-related activities in dACC and LPFC. A large proportion of neurons had activity correlated with δ during post-feedback epochs (**Fig. 6**, referred to as δ-cells, see examples of such cells during late-feedback and inter-trial interval in **Fig. 7A** and **7B**; raster plots and correlation with variable δ can be found in **Supplementary Fig. S7** for the first cell and in **Fig. 9A** for the second cell). Significantly more cells correlated with δ in the dACC than in the LPFC: 252/397 (63%) versus 69/145 (48%; Pearson χ^2 test, 1 df, T = 11.10, $p < 0.001$; **Fig. 6B** and **6C**), which confirmed previous comparisons (Kennerley and Wallis 2009). Consistent with the high learning rate suitable for the task (due to the deterministic reward schedule of the task), the information about the reward prediction error δ from previous trials vanished quickly both in LPFC and dACC compared to other protocols (Seo and Lee 2007). Few dACC cells (31/285, 10.9%) and LPFC cells (9/116, 7.8%) retained a trace of δ from the previous trial in any of the pre-feedback epochs (**Fig. 6B**-**C**). No significant difference was found between dACC and LPFC proportions (Pearson χ^2 test, T = 0.89, $p > 0.3$). Interestingly, only few LPFC δ cells (13/69, 18.8%) revealed a positive correlation (δ^+ cells, *i.e.* neurons responding to unexpected

correct feedback; **Fig. 6B**). The great majority of δ cells in LPFC had negative correlations (56/69, 81.2%), that is, displayed increased activity after errors (δ^- cells; **Fig. 6C**). In comparison, dACC had a higher proportion of δ^+ cells (101/252 δ^+ cells, 40.1%, and 151/202 δ^- cells, 74.8%; see example of such cell in **Fig. 7E**; raster and correlation plots are shown in **Supplementary Fig. S8**). The difference in proportion of δ^+ cells between LPFC and dACC was significant (Pearson χ^2 test, 1 df, T = 10.67, $p < 0.01$). Thus LPFC activity is much more reactive to negative feedback compared to dACC which responds equally to positive and negative feedback.

Previous studies have reported quantitative discrimination of positive reward prediction errors in dACC unit activity (Matsumoto *et al.* 2007; Kennerley and Walton 2011). dACC feedback-related activity might also represent categorical information (i.e. correct, choice error, execution error) rather than quantitative reward prediction errors (Quilodran et al., 2008; see discussion). The present model-based analysis confirms this and also extends it to LPFC feedback-related activity by finding that only very few cells were still correlated with δ when analyzing correct and incorrect trials separately. 10/159 (6.3%) dACC and 2/57 (3.5%) LPFC δ^- cells where still significantly correlated with δ when considering incorrect trials only (multiple regression analysis with bootstrap). These proportions were not significantly different (Pearson χ^2 test, T = 0.62, $p > 0.4$). **Figures 7A** and **7B** illustrate examples of dACC and LPFC neurons which respond to errors without significantly distinguishing between different amplitudes of modeled negative reward prediction errors. 23/101 (22.8%) dACC and 2/13 (15.4%) LPFC δ^+ cells where still significantly correlated with δ on COR trials only. These proportions were not significantly different (Pearson χ^2 test, T = 0.37, $p > 0.5$). **Figure 7E** illustrates the activity of such a cell. In summary, the most striking result regarding feedback-related activity was the differential properties of dACC and LPFC in coding both positive and negative outcomes, LPFC activity being clearly biased toward responding after negative outcomes.

Correlates of outcome uncertainty. Hypotheses on the neural bases of cognitive regulation have been largely inspired by the dynamics of activity variations in dACC and LPFC during behavioral adaptations (Kerns et al. 2004; Brown and Braver 2005). Functions of the dACC are considered to enable monitoring of variations in the history of reinforcements (Seo and Lee 2007, 2008), of the error-likelihood (Brown and Braver 2005), to accordingly adjust behavior. Thus we looked for correlations between single unit activities and the outcome uncertainty U (which progressively increases after elimination of possible targets during search and drops to zero after the first correct trial; see Methods). We observed both positive and negative correlations between dACC neural activity and U (U-cells): 71.8% were positive correlations – higher firing rate during search periods – and 28.2% were negative correlations – higher firing rate during repetition. These proportions are

different from an expected 50%-50% proportion (χ^2 goodness of fit - one sample test, 1 df, $\chi^2 = 39.32$, $p < 0.001$). The population activity of these units correlated with U showed gradual trial-by-trial changes during search, and sharp variations from search to repetition, after the first correct feedback of the problem (see examples of such cells during the post-start epoch in **Fig. 7C**, **D**; see raster and correlation plots in **Supplementary Fig. S7B, C**). These patterns of activity were in opposite direction from changes in reaction times (**Fig. 2B**). They belonged to a larger group of cells that globally discriminated between search and repetition (see a different profile of such type of neurons in the post-target epoch in **Fig. 7F**; see raster and correlation plots in **Supplementary Fig. S8B**). Neural data revealed that U cells were more frequent in dACC (206/397, 52%) than in LPFC (48/145, 33%; Pearson χ^2 test, $T = 15.05$, $p < 0.001$; **Fig. 6D**). Importantly, **Figure 6** shows that, during trials, U was decoded from dACC activity mostly just before and after feedback occurrence. By contrast, U was better decoded during delay (*i.e.,* pre-target epoch) in LPFC. These different dynamics reinforce the idea of an intimate link between U updating and the information provided by feedback for performance monitoring in dACC and, in contrast, of an implication of LPFC in incorporation of U into the decision function in LPFC.

Multiplexed reinforcement-related information. We found that both dACC and LPFC single units multiplexed information about different model variables, with LPFC activity reflecting more integration of information than dACC activity. First, in LPFC the great majority of U-cells (81%, 39/48) were also correlated with one of the model action values while this was true for only 52% (107/206) of dACC U-cells (Pearson χ^2 test, 1 df, $T = 13.68$, $p < 0.001$). Stronger integration was also reflected through higher correlation strengths with multiple variables of the model, as found by a Principal Component Analysis (PCA) on regression coefficients for all dACC and LPFC neurons (**Fig. 8**). The first principal component (PC1) obtained with dACC neurons corresponds in all trial epochs to activity variations mainly related to the outcome uncertainty U and reveals weak links with Q and δ (**Fig. 8A**). In contrast, the two first components (PC1 and PC2) obtained with LPFC neurons both were expressed as a combination of Q and U during pre-feedback epochs (**Fig. 8A**). The PCA also revealed a strong change in the principal components between pre- and post-feedback epochs both in dACC and LPFC and reliably in the two monkeys (**Fig. 8A**), consistent with the post-feedback activity changes and correlations between model variables reported in the previous analyses.

To quantify differences in multiplexing at the single-unit level, we computed an entropy-like index (ELI) of sharpness of encoding of different model variables based on the distributions of correlation strengths between individual cell activities and model variables (see Methods): *e.g.* a neuron with activity correlated with different model variables with similar strengths will have a high ELI; a neuron

with activity highly correlated with only one model variable will have a low ELI (see illustration of different ELI obtained with artificial data illustrating these cases in **Supplementary Fig. S1**). We found a higher ELI in LPFC neurons than in dACC neurons in the pre-touch and pre-feedback epochs (Kruskal-Wallis test, $p < 0.05$) and the opposite effect (i.e. dACC > LPFC) in the early-feedback epoch (Kruskal-Wallis test, $p < 0.05$; **Fig. 8B**). These pre- and post-feedback variations in ELI may reflect different processes: action selection and value updating respectively. Overall, these results reveal higher information integration in LPFC before the feedback, and higher integration in dACC after the feedback.

We then measured the contribution of each model variable to each principal component in each epoch, and combined it with the contribution of each principal component to the global variance in neural activity in each epoch. We deduced a normalized contribution of each model variable to neural activity variance in each epoch (see Methods). Strikingly, in dACC the model variable U dominated (contribution > 50%) in all pre-feedback epochs, while the contribution of δ started increasing in the early-feedback epoch (**Fig. 8C**). In contrast, in LPFC the model variables Q and U had nearly equal contributions to variance during pre-feedback epochs, while the contribution of δ started increasing in the late-feedback epoch, thus later than in dACC. The global entropy in the normalized contributions of model variables to neural activity variance revealed marginally higher in LPFC than in dACC (Kruskal-Wallis test, $p < 0.06$) when analyzed with Model GQLSB2β's variables. These properties of PCA analyses were also true with Model SBnoA2β (**Suppl. Fig. S10**), and the latter effect was found to be even stronger with the latter model (Kruskal-Wallis test, $p < 0.01$; **Suppl. Fig. S10C**), thus confirming the higher information integration in LPFC than in dACC.

Finally, single unit activity could encode different information at different moments in time, corresponding to dynamic coding. More than half LPFC δ-cells (55%, 38/69) – that is, neurons responding to feedback – showed an increase in choice selectivity at the beginning of each new trial in repetition, thus reflecting information about the subsequent choice (see a single cell example in **Fig. 9A**, and a population activity in **Fig. 9C**). In contrast, only 33% (84/252) of dACC δ-cells showed such effect. The difference in proportion between LPFC and dACC was statistically different (Pearson χ^2 test, 1 df, $T = 10.86$, $p < 0.001$; **Fig. 9B**). Thus, while dACC post-feedback activity may mostly be dedicated to feedback monitoring, LPFC activity in response to feedback might reflect the onset of the decision-making process triggered by the outcome.

DISCUSSION

Interaction between performance monitoring and cognitive control hypothetically relies on interactions between dACC and LPFC (e.g. Cohen J.D. et al. 2004). Here we described how the functional link between the two areas might contribute to the regulation of decisions.

In summary, we found that LPFC early-delay activity was more tightly related to monkeys' behavior than dACC activity, displaying higher mutual information with animals' choices than dACC, supporting LPFC's role in action selection. Also, the high choice selectivity in LPFC co-varied with the control level measured from behavior: decreased choice selectivity during the search period, putatively promoting exploration; increased choice selectivity during the repetition period, putatively promoting exploitation. In contrast, this effect was not consistent in dACC. dACC activity correlated with various model variables, keeping track of pertinent information concerning the animal's performance. A calculation of outcome uncertainty (U) correlated with activity changes between exploration and exploitation mostly in dACC, and dominated the contribution to neural activity variance in pre-feedback epochs. Moreover, dACC post-feedback activity appeared earlier than in LPFC and represented positive and negative outcomes with similar proportions while LPFC post-feedback activity mostly tracked negative outcomes.

Reinforcement-related (Q and δ) and task monitoring-related (U) information was multiplexed both in dACC and LPFC, but with higher integration of information before the feedback in LPFC and after the feedback in dACC. LPFC unit activity responding to feedback was also choice selective during early-delay, possibly contributing to decision making, while dACC feedback-related activity – possibly categorizing feedback per se – showed less significant choice selectivity variations. Taken together, these elements suggest that reinforcement-based information and performance monitoring in dACC might participate in regulating decision functions in LPFC.

Mixed information and coordination between areas

Correlations with variables related to reinforcement and actions were found in both structures in accordance with previous studies showing redundancy in information content, although with some quantitative biases (Seo and Lee 2008; Luk and Wallis 2009). However, compared to LPFC, dACC neuronal activity was more selective for outcome uncertainty that could be used to regulate exploration (**Fig. 8**). The PCA analysis showed that multiplexing of reinforcement-related information is stronger in LPFC activity suggesting that this structure receives and integrates these information. In this hypothesis dACC would influence LPFC computations by modulating an action selection process. Such interaction have been interpreted as a motivational or energizing function (from dACC) onto selection mechanisms (in LPFC) (Kouneiher et al. 2009). More specifically, our results support a recently proposed model in which dACC monitors task-relevant signals to compute action values and

keep track of the agent's performance necessary for adjusting behavioral meta-parameters (Khamassi *et al.* 2011; Khamassi *et al.* 2013). In this model, values are transmitted to the LPFC which selects the action to perform. But the selection process (stochastic) is regulated online based on dACC's computations to enable dynamic variations of the control level.

This view preserves the schematic regulatory loop by which performance monitoring acts on cognitive control as proposed by others (Botvinick *et al.* 2001; Cohen J.D. *et al.* 2004). We further suggest a functional structure that reconciles data related to regulatory mechanisms, reinforcement learning, and cognitive control. In particular we point to the potential role of dACC in using reinforcement-related information (such as reward prediction error), relayed through the reward system (Satoh et al. 2003; Enomoto et al. 2011), to regulate global tendencies (formalized by meta-parameters) of adaptation. Interestingly, human dACC (*i.e.,* mid-cingulate cortex) activation co-varies with volatility or variance in rewards and could thereby also participate in regulating learning rates for social or reward-guided behaviors (Behrens *et al.* 2007; Behrens et al. 2009). Kolling and colleagues (Kolling et al. 2012) have recently found that dACC encodes the average value of the foraging environment. This suggests a general involvement of dACC in translating results of performance monitoring and task monitoring into a regulatory level.

The fact that dACC activity correlated with changes in modeled meta-parameters would suggest a general function in the global setting of behavioral strategies. It has been proposed that dACC can be regarded as a filter involved in orienting motor or behavioral commands (Holroyd and Coles 2002), in regulating action decision (Domenech and Dreher 2010), and that it is part of a core network instantiating task-sets (Dosenbach et al. 2006). Interestingly, dACC neural activity encodes specific events that are behaviorally relevant in the context of a task, events that – like the Signal to Change in our task – can contribute to trigger selected adaptive mechanisms (Amiez *et al.* 2005; Quilodran *et al.* 2008). In line with this, Alexander and Brown recently proposed that dACC signals unexpected non-occurrences of predicted outcomes, *i.e.* negative surprise signals, which in their model consist of context-specific predictions and evaluations (Alexander W. H. and Brown 2011). Their model elegantly explains a large amount of reported dACC post-feedback activity. But dACC signals related to positive surprise (Matsumoto *et al.* 2007; Quilodran *et al.* 2008), and to other behaviorally salient events (Amiez *et al.* 2005), suggest an even more general role in processing information useful to guide selected behavioral adaptations.

Exploration

Following a standard reinforcement learning framework, exploratory behavior was here associated to low β values, which flatten the probability distribution of competing actions in models and simulations (Khamassi *et al.* 2011). Although the precise molecular and cellular mechanisms underlying shifts between exploration and exploitation are not yet known, accumulating evidence suggest that differential levels of activation of D1 and D2 dopamine receptors in the prefrontal cortex may produce distinct states of activity: a first state allowing multiple network representations nearly simultaneously and thus permitting "an exploration of the input space"; a second state where the influence of weak inputs on PFC networks is shut off so as to stabilize one or a limited set of representations, which would then have complete control on PFC output and thus promote exploitation (Durstewitz and Seamans 2008). The consistent variations of LPFC choice selectivity between search and repetition periods suggest that such mechanism could also underlie exploration during behavioral adaptation.

However, this should not be interpreted as an assumption that monkeys' behavior is purely random during search periods of the task (see model-based analysis of behavior). In fact, animals often display structured and organized exploratory behaviors as also revealed by our behavioral analyses. For instance, when facing a new open arena, rodents display sequential stages of exploration, first remaining around the nest position, second moving along walls and third visiting the center of the arena (Fonio et al. 2009). Non-human primates also use exploration strategies, such as optimized search trajectories adapted to the search space configuration (De Lillo et al. 1997), trajectories that can evolve based on reinforcement history along repeated exposure to the same environment (Desrochers et al. 2010). In ecological large scale environments search strategies are best described by correlated random or Levy walks and are modulated by various environmental parameters (Bartumeus et al. 2005).

One possible interpretation of our results is that decreases of choice selectivity in LPFC during search could reduce the amount of information about choice and ergo release biases in the influence on downstream structures such as the basal ganglia. In this way, efferent structures could express their own exploratory decisions. Consistent with this, it has been recently suggested that variations of tonic dopamine in the basal ganglia could also affect the exploration-exploitation trade off in decision-making (Humphries et al. 2012).

The prefrontal cortex might also contribute to the regulation of exploration based on current uncertainty (Daw N. D. *et al.* 2006; Frank *et al.* 2009). Uncertainty-based control could bias decision towards actions that provide very variable quantities of reward so as to gain novel information and reduce uncertainty. In our task, outcome uncertainty variations – progressive increase during search and drop to zero during repetition – can be confounded with other similar performance monitoring

measures such as the feedback history (Khamassi *et al.* 2011) or variations of attentional level. Nevertheless, they co-varied with the animal's reaction times and were mostly encoded by dACC neurons, thus revealing a possible relevance of this information for behavioral control in our task. It should be noted that outcome uncertainty is distinct from action uncertainty which would be confounded in our task with other task monitoring variables such as conflict (Botvinick *et al.* 2001) and error-likelihood (Brown and Braver 2005). All of them gradually and monotonically decrease along a typical problem of the PS task and remain low during repetition. We found neurons with such activity profile (e.g. **Fig. 7F**), however in about half the proportion of U-cells. More work is required to understand whether these different task monitoring measures are distributed and coordinated within the dACC-LPFC system.

Reinforcement learning or working memory?

It has been recently suggested that model-based investigations of adaptive mechanisms often mix and confound reinforcement learning mechanisms and working memory updating (Collins and Frank, 2012). In particular, rapid improvements in behavioral performance during decision-making tasks can be best explained by gating mechanisms in computational models of the prefrontal cortex rather than by slow adaptation usually associated with dopamine-dependent plasticity in the basal ganglia. In the present study, the fact that Models SBnoA and SBnoA2β (with a high learning rate α fixed to 1) and Models GQLSB and GQLSB2β (where α is a free-metaparameter between 0 and 1) produce a non-different fitting score on monkey behavior suggests that behavior in this task might fall into such a case. Under this interpretation, rapid behavioral adaptations would rely on gating appropriate flows of information between dACC and LPFC. In fact, the increase of LPFC activity mostly after negative and not positive outcomes, and the interaction with spatial selectivity, might reflect gating working memory or planning processes at the time of adaptation, rather than direct outcome-related responses. An alternative hypothesis that cannot be excluded is that in this type of deterministic task animals still partly rely on reinforcement learning mechanisms, but would progressively learn to employ a high learning rate during the long pretraining phase. The fact that a group of behavioral sessions were better fitted with α between 0.3 and 0.9 when α was not fixed to 1 (i.e. in Model GQLSB; **Supplementary Fig. S2C**) reveals a continuum in the range of optimized α values which could be the result of a progressive but incomplete increase of the learning rate during pretraining. Such adaptation in rate might have also contributed to the weak quantitative coding of reward prediction errors. Further investigations will be required to answer this question, in particular by precisely characterizing monkey behavioral performance during the pretraining phase and the associated changes in information coding in prefrontal cortical regions.

Network regulation and decisions in LPFC

We reported new data on the possible functional link between LPFC and dACC. However, we have no evaluation of putative dynamical and direct interactions between neurons of the two regions. Functional coordination of local field potentials between LPFC and dACC has been described but evidence for direct interactions is scarce (Rothe *et al.* 2011). The schematized modulatory function from dACC performance monitoring into LPFC decision process could in fact be indirect. For instance, it has been proposed that norepinephrine instantiates gain (excitability) variations in LPFC, and that this mechanism would be regulated by dACC afferences to the locus coeruleus (Aston-Jones and Cohen 2005; Cohen J. D. *et al.* 2007). Average activity variations in dACC and LPFC observed in our recordings could be a consequence of such activity gain changes. Gain modulation and biased competition are two mechanisms by which attentional signals can operate (Wang 2010). Increased working memory load, higher cognitive control, or attentional selection are concepts widely used to interpret prefrontal activity modulations dependent on task requirements (Miller and Cohen 2001; Leung et al. 2002; Kerns *et al.* 2004). Note that these concepts are closely related and have similar operational definitions (Barkley 2001; Miller and Cohen 2001; Cohen J.D. *et al.* 2004).

Recently, Kaping and colleagues have shown that spatial attentional and reward valuation signals are observed in different subdivisions of the fronto-cingulate region (Kaping *et al.* 2011). Correlates of spatial attention selectivity were found in both dACC and LPFC, together with correlates of valuation, and independently of action plans. These signals would contribute to top-down attentional control of information (Kaping *et al.* 2011). Here we also verified that values were coded independently of choices by showing significant correlation with Q-values even after exclusion of trials selecting the neuron's preferred target.

The present study revealed two effects of task periods on frontal activity that would reflect variations in control and decision: an increased average firing rate and changes in recruited neural populations during exploration in both dACC and LPFC, and an increased spatial selectivity in LPFC during repetition. The latter would argue against a reduction of control implemented by LPFC during repetition. This however suggests that transitions between exploration and repetition involve a complex interplay between global unselective regulations and refined selection functions, and that qualitative changes in control occurred between search and repetition.

Finally, studies in rodents suggest that adaptive changes in behavioral strategies are also accompanied by global dynamical state transitions of prefrontal activity (Durstewitz et al. 2010). Our analyses showed that for both LPFC and dACC the neural populations participating in exploratory versus exploitative periods of the task differ significantly. We have also previously shown that the

oscillatory coordination between the two areas changes from one period to the other (Rothe *et al.* 2011). Hence, a dynamical system perspective might be imperative to explain cognitive flexibility and its neurobiological substrate with more precision.

References

Alexander WH, Brown JW. 2010. Computational Models of Performance Monitoring and Cognitive Control. Topics in Cognitive Science. 2: 658-677.

Alexander WH, Brown JW. 2011. Medial prefrontal cortex as an action-outcome predictor. Nat Neurosci. 14: 1338-1344.

Amiez C, Joseph JP, Procyk E. 2005. Anterior cingulate error-related activity is modulated by predicted reward. Eur J Neurosci. 21: 3447-3452.

Amiez C, Neveu R, Warrot D, Petrides M, Knoblauch K, Procyk E. 2013. The location of feedback-related activity in the midcingulate cortex is predicted by local morphology. J Neurosci. 33: 2217-2228.

Aston-Jones G, Cohen JD. 2005. An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. Annu Rev Neurosci. 28: 403-450.

Barkley RA. 2001. Linkages between attention and executive functions. In: Reid Lyon G, Krasnegor NA, eds. Attention, memory and executive function P.H. Brooks p 307-326.

Barraclough DJ, Conroy ML, Lee D. 2004. Prefrontal cortex and decision making in a mixed-strategy game. Nat Neurosci. 7: 404-410.

Bartumeus F, da Luz MG, Viswanathan GM, Catalan J. 2005. Animal search strategies: a quantitative random-walk analysis. Ecology. 86: 3078-2087.

Behrens TE, Hunt LT, Rushworth MF. 2009. The computation of social behavior. Science. 324: 1160-1164.

Behrens TE, Woolrich MW, Walton ME, Rushworth MF. 2007. Learning the value of information in an uncertain world. Nat Neurosci. 10: 1214-1221.

Botvinick MM, Braver TS, Barch DM, Carter CS, Cohen JD. 2001. Conflict monitoring and cognitive control. Psychol Rev. 108: 624-652.

Brown JW, Braver TS. 2005. Learned predictions of error likelihood in the anterior cingulate cortex. Science. 307: 1118-1121.

Cohen JD, Aston-Jones G, Gilzenrat MS. 2004. A systems-level perspective on attention and cognitive control. In: Posner MI, ed. Cognitive Neuroscience of attention New York: Guilford p 71-90.

Cohen JD, McClure SM, Yu AJ. 2007. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. Philos Trans R Soc Lond B Biol Sci. 362: 933-942.

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

Daw ND. 2011. Trial-by-trial data analysis using computational models. In: Affect, Learning and Decision Making. New York: Oxford University Press

Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ. 2006. Cortical substrates for exploratory decisions in humans. Nature. 441: 876-879.

De Lillo C, Visalerberghi E, Aversano M. 1997. The organization of exhaustive searches in a patchy space by capuchin monkeys (cebus apella). J Comp Psychol. 111.

Dehaene S, Kerszberg M, Changeux JP. 1998. A neuronal model of a global workspace in effortful cognitive tasks. Proc Natl Acad Sci U S A. 95: 14529-14534.

Desrochers TM, Jin DZ, Goodman ND, Graybiel AM. 2010. Optimal habits can develop spontaneously through sensitivity to local cost. Proc Natl Acad Sci U S A. 107: 20512-20517.

Domenech P, Dreher JC. 2010. Decision threshold modulation in the human brain. J Neurosci. 30: 14305-14317.

Dosenbach NU, Visscher KM, Palmer ED, Miezin FM, Wenger KK, Kang HC, Burgund ED, Grimes AL, Schlaggar BL, Petersen SE. 2006. A core system for the implementation of task sets. Neuron. 50: 799-812.

Doya K. 2002. Metalearning and neuromodulation. Neural Netw. 15: 495-506.

Durstewitz D, Seamans JK. 2008. The dual-state theory of prefrontal cortex dopamine function with relevance to catechol-o-methyltransferase genotypes and schizophrenia. Biol Psychiatry. 64: 739-749.

Durstewitz D, Vittoz NM, Floresco SB, Seamans JK. 2010. Abrupt transitions between prefrontal neural ensemble states accompany behavioral transitions during rule learning. Neuron. 66: 438-448.

Enomoto K, Matsumoto N, Nakai S, Satoh T, Sato TK, Ueda Y, Inokawa H, Haruno M, Kimura M. 2011. Dopamine neurons learn to encode the long-term value of multiple future rewards. Proc Natl Acad Sci U S A. 108: 15462-15467.

Fonio E, Benjamini Y, Golani I. 2009. Freedom of movement and the stability of its unfolding in free exploration of mice. Proc Natl Acad Sci U S A. 106: 21335-21340.

Frank MJ, Doll BB, Oas-Terpstra J, Moreno F. 2009. Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nat Neurosci. 12: 1062-1068.

Holroyd CB, Coles MG. 2002. The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. Psychol Rev. 109: 679-709.

Humphries MD, Khamassi M, Gurney K. 2012. Dopaminergic Control of the Exploration-Exploitation Trade-Off via the Basal Ganglia. Frontiers in neuroscience. 6: 9.

Ishii S, Yoshida W, Yoshimoto J. 2002. Control of exploitation-exploration meta-parameter in reinforcement learning. Neural Netw. 15: 665-687.

Ito M, Doya K. 2009. Validation of decision-making models and analysis of decision variables in the rat basal ganglia. J Neurosci. 29: 9861-9874.

Kaping D, Vinck M, Hutchison RM, Everling S, Womelsdorf T. 2011. Specific contributions of ventromedial, anterior cingulate, and lateral prefrontal cortex for attentional selection and stimulus valuation. PLoS Biol. 9: e1001224.

Kennerley SW, Wallis JD. 2009. Evaluating choices by single neurons in the frontal lobe: outcome value encoded across multiple decision variables. Eur J Neurosci. 29: 2061-2073.

Kennerley SW, Walton ME. 2011. Decision making and reward in frontal cortex: complementary evidence from neurophysiological and neuropsychological studies. Behav Neurosci. 125: 297-317.

Kennerley SW, Walton ME, Behrens TE, Buckley MJ, Rushworth MF. 2006. Optimal decision making and the anterior cingulate cortex. Nat Neurosci. 9: 940-947.

Kerns JG, Cohen JD, MacDonald AW, 3rd, Cho RY, Stenger VA, Carter CS. 2004. Anterior cingulate conflict monitoring and adjustments in control. Science. 303: 1023-1026.

Khamassi M, Enel P, Dominey PF, Procyk E. 2013. Medial prefrontal cortex and the adaptive regulation of reinforcement learning parameters. Prog Brain Res. 202: 441-464.

Khamassi M, Lallee S, Enel P, Procyk E, Dominey PF. 2011. Robot cognitive control with a neurophysiologically inspired reinforcement learning model. Front Neurorobot. 5: 1.

Kolling N, Behrens TE, Mars RB, Rushworth MF. 2012. Neural mechanisms of foraging. Science. 336: 95-98.

Kouneiher F, Charron S, Koechlin E. 2009. Motivation and cognitive control in the human prefrontal cortex. Nat Neurosci. 12: 939-945.

Krichmar JL. 2008. The neuromodulatory system – a framework for survival and adaptive behavior in a challenging world. Adapt Behav. 16: 385-399.

Landmann C, Dehaene S, Pappata S, Jobert A, Bottlaender M, Roumenov D, Le Bihan D. 2007. Dynamics of prefrontal and cingulate activity during a reward-based logical deduction task. Cereb Cortex. 17: 749-759.

Lau B. 2014. Matlab code for diagnosing collinearity in a regression design matrix. figshare. http://dx.doi.org/10.6084/m9.figshare.1008225.

Leung HC, Gore JC, Goldman-Rakic PS. 2002. Sustained mnemonic response in the human middle frontal gyrus during on-line storage of spatial memoranda. J Cogn Neurosci. 14: 659-671.

Luk CH, Wallis JD. 2009. Dynamic encoding of responses and outcomes by neurons in medial prefrontal cortex. J Neurosci. 29: 7526-7539.

MacDonald AW, 3rd, Cohen JD, Stenger VA, Carter CS. 2000. Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. Science. 288: 1835-1838.

Matsumoto M, Matsumoto K, Abe H, Tanaka K. 2007. Medial prefrontal cell activity signaling prediction errors of action values. Nat Neurosci. 10: 647-656.

McClure SM, Gilzenrat MS, Cohen JD. 2006. An exploration–exploitation model based on norepinephrine and dopamine activity. In: Weiss Y, Sholkopf B, Platt J, eds. Advances in neural information processing systems MIT Press, Cambridge, MA p 867–874.

Miller EK, Cohen JD. 2001. An integrative theory of prefrontal cortex function. Annu Rev Neurosci. 24: 167-202.

Panzeri S, Senatore R, Montemurro MA, Petersen RS. 2007. Correcting for the sampling bias problem in spike train information measures. J Neurophysiol. 98: 1064-1072.

Panzeri S, Treves A. 1996. Analytical estimates of limited sampling biases in different information measures. Network: Computation in Neural Systems. 7: 87-107.

Procyk E, Goldman-Rakic PS. 2006. Modulation of dorsolateral prefrontal delay activity during self-organized behavior. J Neurosci. 26: 11313-11323.

Procyk E, Tanaka YL, Joseph JP. 2000. Anterior cingulate activity during routine and non-routine sequential behaviors in macaques. Nat Neurosci. 3: 502-508.

Quian Quiroga R, Panzeri S. 2009. Extracting information from neuronal populations: information theory and decoding approaches. Nat Rev Neurosci. 10: 173-185.

Quilodran R, Rothé M, Procyk E. 2008. Behavioral shifts and action valuation in the anterior cingulate cortex. Neuron. 57(2): 314–325.

Rothe M, Quilodran R, Sallet J, Procyk E. 2011. Coordination of High Gamma Activity in Anterior Cingulate and Lateral Prefrontal Cortical Areas during Adaptation. J Neurosci. 31: 11110-11117.

Rushworth MF, Behrens TE. 2008. Choice, uncertainty and value in prefrontal and cingulate cortex. Nat Neurosci. 11: 389-397.

Satoh T, Nakai S, Sato T, Kimura M. 2003. Correlated coding of motivation and outcome of decision by dopamine neurons. J Neurosci. 23: 9913-9923.

Schultz W, Dayan P, Montague PR. 1997. A neural substrate of prediction and reward. Science. 275: 1593-1599.

Schweighofer N, Doya K. 2003. Meta-learning in reinforcement learning. Neural Netw. 16: 5-9.

Seo H, Lee D. 2007. Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game. J Neurosci. 27: 8366-8377.

Seo H, Lee D. 2008. Cortical mechanisms for reinforcement learning in competitive games. Philos Trans R Soc Lond B Biol Sci. 363: 3845-3857.

Seo H, Lee D. 2009. Behavioral and neural changes after gains and losses of conditioned reinforcers. J Neurosci. 29: 3627-3641.

Sutton RS, Barto AG. 1998. Reinforcement learning: an introduction. Cambridge, MA London, England: MIT Press.

Treves A, Panzeri S. 1995. The upward bias in measures of information derived from limited data samples. Neural Comput. 7: 399-407.

Vogt BA, Vogt L, Farber NB, Bush G. 2005. Architecture and neurocytology of monkey cingulate gyrus. J Comp Neurol. 485: 218-239.

Wang XJ. 2010. Neurophysiological and computational principles of cortical rhythms in cognition. Physiol Rev. 90: 1195-1268.

Wilson CR, Gaffan D, Browning PG, Baxter MG. 2010. Functional localization within the prefrontal cortex: missing the forest for the trees? Trends Neurosci. 33: 533-540.

Acknowledgments

The authors would like to thank Jacques Droulez, Mark D. Humphries, Henry Kennedy, Olivier Sigaud and Charlie R.E. Wilson for comments on an early version of the manuscript, and Francesco P. Battaglia and Erika Cerasti for useful discussions. They also would like to thank anonymous reviewers for thorough comments and questions which helped drastically improve the manuscript. This work was supported by the Agence Nationale de la Recherche ANR LU2 and EXENET, Région Rhône-Alpes projet Cible, and by the labex CORTEX ANR-11-LABX-0042 for EP; by EU FP7 Project Organic (ICT 231267) for PFD; By Facultad de Medicina Universidad de Valparaíso (MECESUP UVA-106) and by Fondation pour la Recherche Médicale for RQ; By ANR (Amorces and Comprendre) for PFD and MK.

Table 1

Score obtained by each tested theoretical model, models' characteristics, and model performances to fit monkey choices for Optimization (Opt) and Test sessions.

Models	*r* [1]	*RL*[2]	N_P [3]	*Opt -LL*[4]	*Opt NL*[5]	*Opt %*[6]	*Opt -LPP*[7]	*Opt BIC/2*[8]	*Opt AIC/2*[9]	*Test -LL*[4]	*Test NL*[5]	*Test %*[6]
GQLSB2β	Y	Y	4	3290	.5921	83.47	3459	3360	3298	29732	.5830	74.17
SBnoA2β	Y	N	3	3385	.5831	84.13	3422	3438	3391	30901	.5708	73.11
GQLSB	Y	Y	3	3355	.5859	83.80	3502	3408	3361	29539	.5850	73.43
SBnoA	Y	N	2	3454	.5768	84.29	3480	3489	3458	30613	.5738	72.59
SBnoF	Y	N	1	3586	.5648	84.43	3604	3604	3588	32169	.5578	71.61
GQLBnoS	Y	Y	3	3721	.5528	78.59	3847	3773	3727	33274	.5467	69.47
GQLSnoB	Y	Y	3	3712	.5536	76.66	3843	3764	3718	31501	.5646	70.12
GQLnoSnoB	Y	Y	3	4253	.5079	69.14	4292	4305	4259	35376	.5262	66.60
GQL	N	Y	3	5590	.4104	65.10	5994	5643	5596	49282	.4089	53.20
QL	N	Y	2	5960	.3869	44.92	7755	5995	5964	59734	.3382	48.78
ClockS	Y	N	2	5249	.4333	70.92	5841	5284	5253	47504	.4223	58.71
RandS	Y	N	1	4607	.4800	69.43	4621	4624	4609	39488	.4884	63.73

[1] Resetting action values at the beginning of each new problem (Yes or No)
[2] Reinforcement Learning (RL) mechanisms or not
[3] Number of free meta-parameters
[4] Negative Log Likelihood
[5] Normalized Likelihood over all trials
[6] Percentage of trials where the model correctly predicted monkey choice
[7] Log of Posterior Probability
[8] Bayesian Information Criterion
[9] Akaike Information Criterion

FIGURE LEGENDS

Figure 1. Task, recording sites, and trial epochs for analyses. (**A**) Problem Solving task. Monkeys had to find by trial and error which target, presented in a set of four, was rewarded. Trial: description of events in a trial (see methods). A juice reward is delivered if the trial was correct while only a blank screen is presented for errors. Problem: In each trial the animal could select a target until the solution was discovered (search period). Each block of trials (or problem) contained a search period and a repetition period during which the correct response was repeated at least three times. A Signal to Change (SC) is presented on the screen to indicate the beginning of a new problem. (**B**) Recording sites for LPFC (grey spots) and dACC (black spots) for the two monkeys. dACC recordings covered a region in the dorsal bank of the anterior cingulate sulcus, at stereotaxic levels superior to A+30, i.e. rostral levels of the mid-cingulate cortex. Recording sites in LPFC were located on the posterior third of the principal sulcus. (**C**) Target identifications and definition of epochs used for single unit analyses.

Figure 2. Model-based behavioral analyses. (**A-left**) Illustration of the trial by trial evolution of action values after meta-parameters optimization so that the model behaves similarly to the monkey. Sample data presented for 100 successive trials. The barcode on the top indicates the current correct target. Each of the 4 targets is associated to one grey level. Head arrows represent the Signal to Change (SC) presented at the beginning of each new problem. The second barcode indicates the target selected by the animal in each trial. The third barcode indicates the target selected by the model based on the feedback obtained by the animal. Variation of action values for each of the 4 targets are represented by curves. The high learning rate (α=0.9) that resulted from the optimization produced sharp variations of action values. The data are presented for two models (SBnoA and GQLSB). (**A-middle**) Proportion of shifts after SC for monkeys M and P. (**A-right**) Proportion of selection of each target in the first trial of each problem across sessions of recordings. Each line represents one target position. (**B**) Reaction times (RT) measured in two monkeys averaged for typical optimal problems: those where the monkey made 2 errors (INC1 and INC2) during the search period, found the correct target (CO1) in the third trial, and repeated the correct choice from 3 to 7 times (CO2 to CO8), depending on the problem's length, during repetition trials. **: $p<0.005$, ***:$p<0.001$. (**C**) Percentage of errors made by the animal during the repetition periods against the exploration rate β_R of the repetition periods. One data point per session. (**D**) Scores obtained by each tested model during the model comparison analysis (see methods). Opt -LL = negative log-likelihood on the optimization dataset. -LPP = negative log of posterior probability. BIC = Bayesian Information Criterion. AIC = Akaike Information Criterion. Test -LL = negative log-likelihood on the test dataset. (**E**)

Distribution of exploration meta-parameters obtained after optimization of the model on monkey's behavior using distinct degrees of freedom during the search period (β_S) and the repetition period (β_R).

Figure 3. Variations of early-delay activity and choice selectivity. (**A-top**) Proportions of dACC and LPFC cells with a higher activity during search or repetition. (**A-bottom**) Proportions of dACC and LPFC cells with a higher choice selectivity during Sea or Rep. (**B**) Number of cells with significant changes (in grey) in average unit activity between search (Sea) and repetition (Rep). The histograms represent the distribution of indices of variation of activity from search to repetition computed in the early-delay epoch with equation (9) in dACC and LPFC neurons. Grey bars represent neurons with significantly different activity between search and repetition trials (Kruskal-Wallis test, $p < 0.05$). White bars represent neurons with non-significantly different activity in search and repetition. (**C**) Increase of choice selectivity from search to repetition in the two structures. Stars indicate statistically significant comparisons *: $p<0.05$, **: $p<0.01$. (**D**) Compared to dACC neurons (grey bars), a higher proportion of LPFC neurons showed significant mutual information between the early-delay average firing rate and the animal's choice. Dashed grey and black lines represent the medians for dACC and LPFC respectively.

Figure 4. Early-delay choice selectivity varies with exploration level. (**A**). The average choice selectivity index is presented for units recorded in dACC (top) and LPFC (bottom), in sessions grouped according to the fitted model's exploration meta-parameters for repetition (β_R). The average population index is measured for search (grey bars) and repetition (white bars) trials in the early-delay epoch, separately for sessions where β_R was inferior or superior to 5. Stars indicate statistically significant comparisons. *: $p<0.05$. (**B**). Proportion of dACC and LPFC early-delay choice selective neurons during repetition periods of sessions where βR was small (<5) or large (>5). Only LPFC revealed a significant change in proportion.

Figure 5. Two examples of action value neurons. (**A**) dACC unit negatively correlated with the value of target #4. (Top) plot of single trial activity (black dots) measured in the post-target epoch against the Q-value, for trials where the animal chose target 4. Large grey dots represent the average for one decile of the value distribution and are just used for illustration. The dashed line represents the linear regression computed from single trial data. (Middle) peri-stimulus histograms aligned on target onset (Target ON) and the corresponding raster plots for trials in which the animal chose target 4 (in black) and for the other trials (in grey). The post-target epoch is represented in grey on the time line.

(Bottom) trial by trial evolution of the average activity measured in the post-target epoch during successive trials in a session. The upper grey barcode represents the correct target to be chosen (4 greys for 4 target positions; corresponding target number is indicated above the bar code). The second barcode represents the target chosen by the animal in each trial. Below, the graph represents the average activity for each trial and, the trial by trial evolution of key model variables. Grey areas represent trials where the animal selected target #4. See main text for details. (**B**) LPFC neuron with a positive correlation with the value of target #2 during the post-target epoch. Conventions as in A.

Figure 6. Proportions of dACC and LPFC cells with activity correlated with one of the model variables (Q, δ, and U) in one of the 9 trial epochs (bars from left to right: pre-start, delay, pre-target, post-target, pre-touch, pre-feedback, early-feedback, late-feedback, ITI). The white and black arrow heads indicate touch and feedback respectively. There were more LPFC cells correlated with one of the action-values (Q, in **A**). In **B** and **C**, δ+ or δ- represent respectively positive and negative correlations with δ. A higher proportion of dACC cells were either positively or negatively correlated with δ (δ+ or δ-) compared to LPFC. These cells mostly responded during post-feedback epochs, and very few cells retained a trace of the previous δ during the beginning of the next trial (pre-feedback epochs). There were more U cells in dACC than in LPFC (in **D**). See text for details. **E**. Proportion of cells, for each epoch, showing a significant correlation with at least one model variable.

Figure 7. Six examples (A-F) of unit activity correlated with some of the model's variables. Line graphs represent average activity aligned on feedback (FB), trial start, or target onset. The grey intensity of lines corresponds to the different trial types as described in the bar graphs below. The grey zone on each time axis represents the epoch used for average measures displayed in the bar graph. Bar graphs represent, for each unit, the average activity measured in the time epoch for the 6 trial types of a typical problem. The trial types in search are: sea1 (first error trial, black), sea2 (second error trial, dark grey), sea3 (third trial in search for activity measured before feedback, grey), and CO1 (first correct trial for activity measured after the feedback, grey in A, B, and E). Trial types in repetition are CO2, CO3, and CO4 (light grey). (**A**) example of dACC activity negatively correlated with RPE (δ-). (**B**) example of LPFC activity negatively correlated with RPE (δ-). (**C**) example of LPFC activity correlated with U. (**D**) example of dACC activity negatively correlated with U. (**E**) example of dACC activity positively correlated with RPE (δ+). (**F**) example of activity discriminating search and repetition but with a different profile than U.

Figure 8. Multiplexing of information and variations during epochs in dACC and LPFC. (**A**) A principal component analysis was performed on the regression coefficients found for each neuron and for each model variable (Q: the action value of the animal's preferred target, δ, and U; Model GQLSB2β). The absolute value of the eigen values for each principal component computed during the early-feedback epoch are shown in each matrix for one trial epoch. Black denotes strong weights. Data are presented for each monkey M and P. (**B**) Evolution of the entropy-like factor on regression coefficients computed for 2 variables Q and U, and Q and δ. A * indicates a statistically significant difference between dACC (in grey) and LPFC (in black). (**C**) Proportion of total variance explained by each model variable over the 3 PCs for dACC and LPFC data along trial epochs. See main text for details.

Figure 9. Variations of choice selectivity in δ-cells. (**A**) Example of a LPFC cell responding after errors (activity negatively correlated with δ in the late-feedback epoch) and showing an increase in choice selectivity at the beginning of trials. Left: error trials are illustrated in grey, correct trials in black. Right: trials are grouped by chosen targets. 4 grey curves for 4 target locations. (**B**) Percentage of dACC and LPFC δ-cells showing a significant increase in choice selectivity from search to repetition. (**C**) Averaged population activity (50 ms bins) of all dACC (left) and LPFC (right) units negatively correlated with δ. For each cell, the activity was averaged separately for trials in which the animal selected the cell's preferred target (black plain line), the second preferred target (black dashed line), the third (gray dashed line) or the least preferred target (gray plain line). The activity is represented in 3s windows centered on the feedback time (FB, Left) and on the next trial start (ST, Right), for search trials (Top) and repetition trials (Bottom). In LPFC, negative δ cells showed an increase in choice selectivity in the post-start epoch of repetition trials.

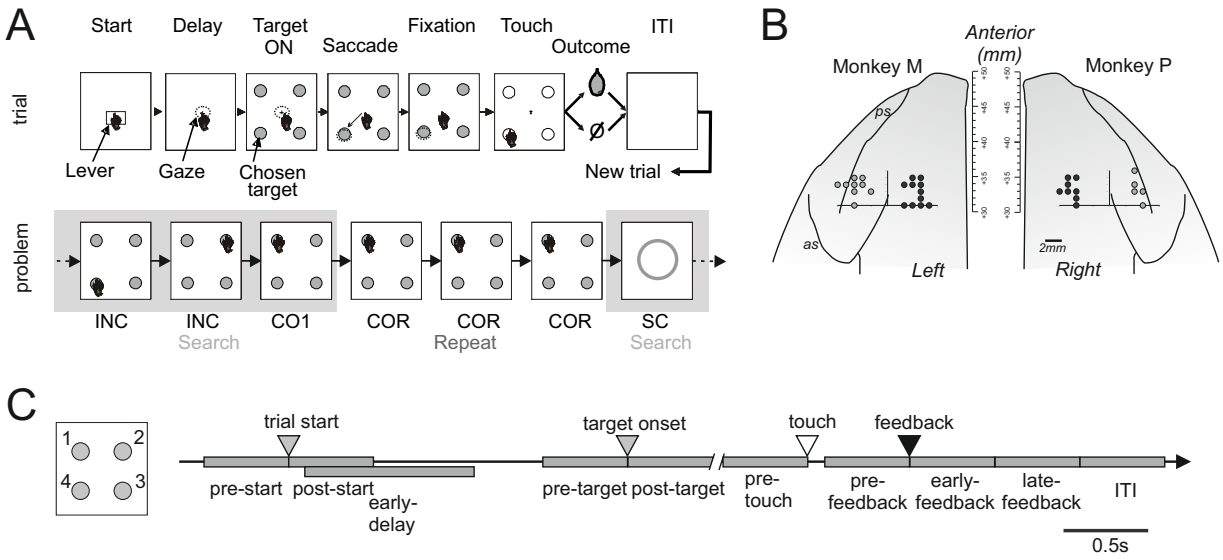

A
Start
Delay
Target ON
Saccade
Fixation
Touch
Outcome
ITI
trial
Lever
Gaze
Chosen target
New trial
problem
INC
INC
Search
CO1
COR
COR
Repeat
COR
SC
Search
B
Anterior (mm)
Monkey M
Monkey P
ps
as
2mm
Left
Right
C
1
2
3
4
trial start
target onset
touch
feedback
pre-start
post-start
early-delay
pre-target
post-target
pre-touch
pre-feedback
early-feedback
late-feedback
ITI
0.5s

Figure 1

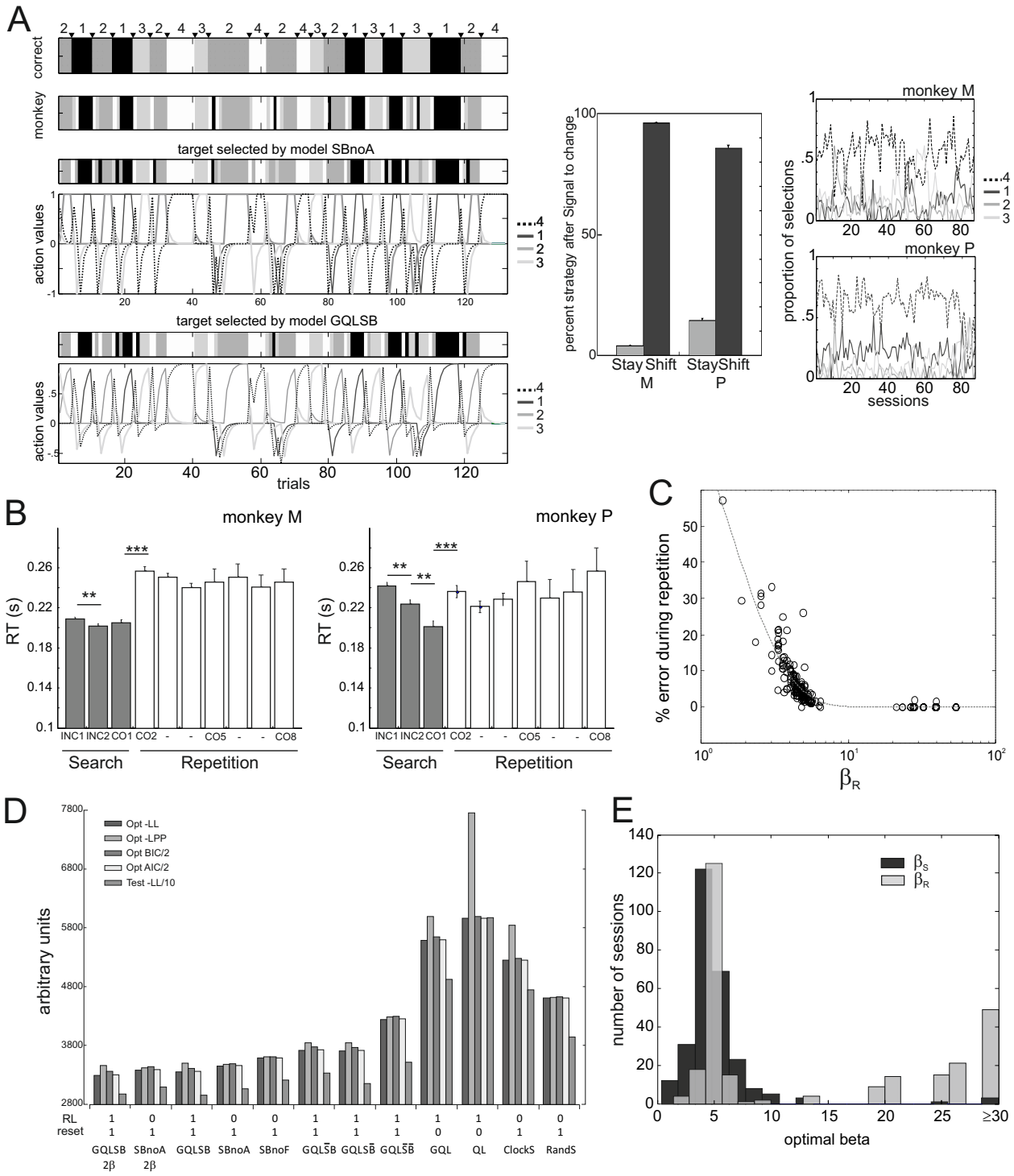

A
correct
monkey
target selected by model SBnoA
action values
target selected by model GQLSB
action values
trials
percent strategy after Signal to change
Stay Shift Stay Shift
M P
monkey M
monkey P
proportion of selections
sessions
B
monkey M
monkey P
RT (s)
INC1 INC2 CO1 CO2 CO5 CO8
Search
Repetition
C
% error during repetition
β_R
D
arbitrary units
Opt -LL
Opt -LPP
Opt BIC/2
Opt AIC/2
Test -LL/10
RL
reset
GQLSB 2β
SBnoA 2β
GQLSB
SBnoA
SBnoF
GQLS̄B
GQLSB̄
GQLS̄B̄
GQL
QL
ClockS
RandS
E
number of sessions
β_S
β_R
optimal beta

Figure 2

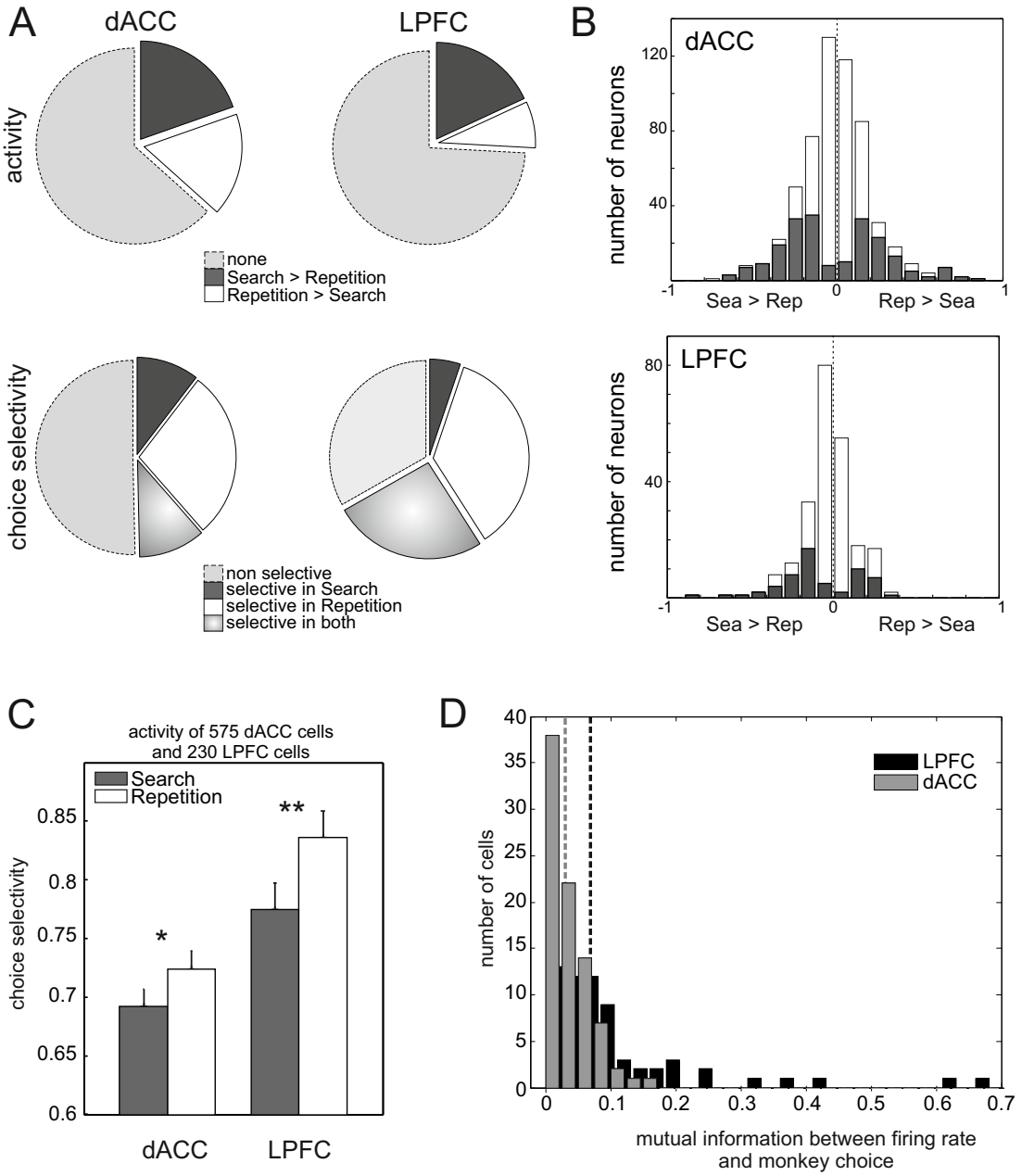

A
dACC
LPFC
activity
none
Search > Repetition
Repetition > Search
choice selectivity
non selective
selective in Search
selective in Repetition
selective in both
B
dACC
number of neurons
120
80
40
-1
0
1
Sea > Rep
Rep > Sea
LPFC
80
40
C
activity of 575 dACC cells
and 230 LPFC cells
Search
Repetition
choice selectivity
0.85
0.8
0.75
0.7
0.65
0.6
*
**
dACC
LPFC
D
LPFC
dACC
number of cells
40
35
30
25
20
15
10
5
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
mutual information between firing rate
and monkey choice

Figure 3

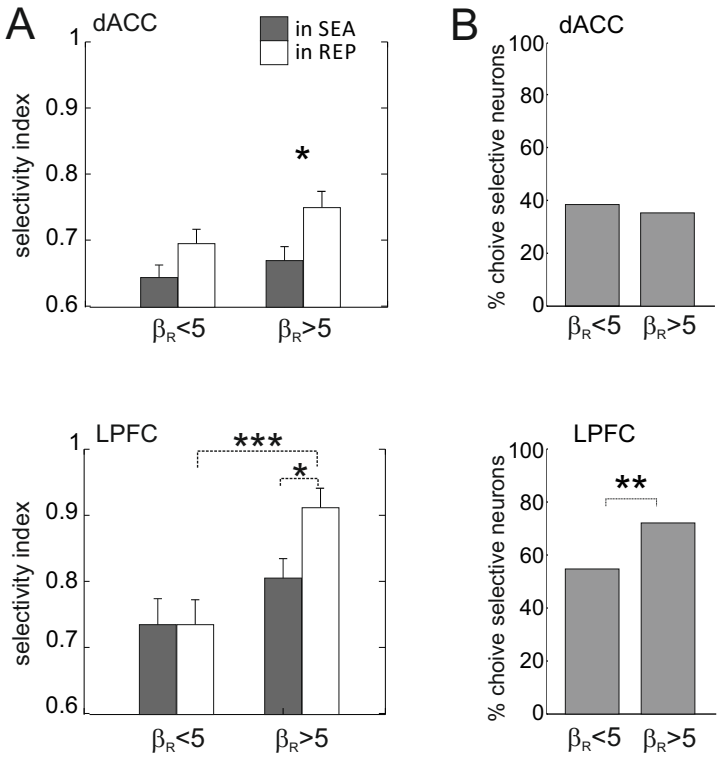

A
dACC
in SEA
in REP
selectivity index
1
0.9
0.8
0.7
0.6
*
β_R<5
β_R>5
B
dACC
% choive selective neurons
100
80
60
40
20
0
β_R<5
β_R>5
LPFC
***
*
selectivity index
1
0.9
0.8
0.7
0.6
β_R<5
β_R>5
LPFC
**
% choive selective neurons
100
80
60
40
20
0
β_R<5
β_R>5

Figure 4

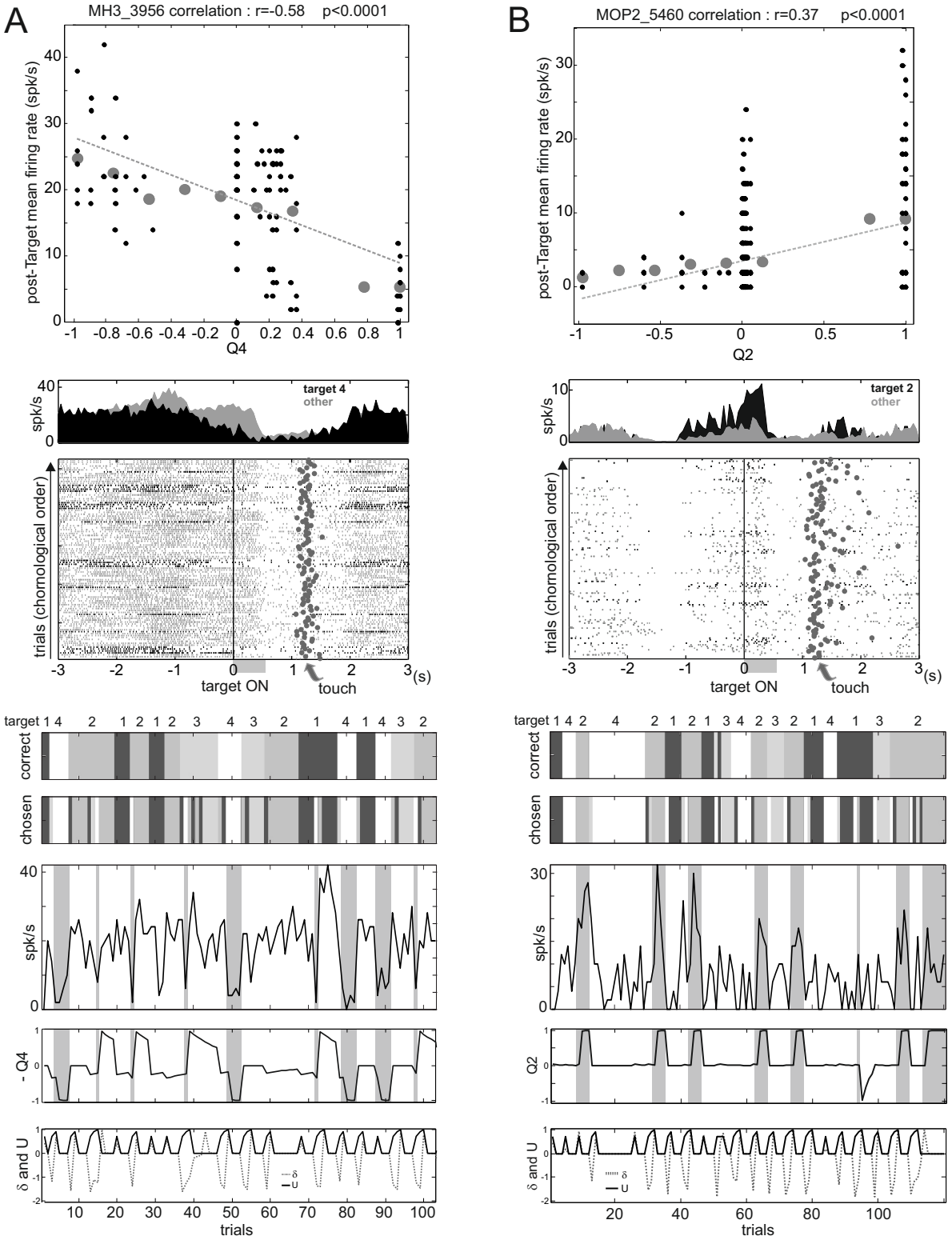

A
MH3_3956 correlation : r=-0.58 p<0.0001
post-Target mean firing rate (spk/s)
Q4
target 4
other
spk/s
trials (chronological order)
target ON
touch
(s)
target 1 4 2 1 2 1 2 3 4 3 2 1 4 1 4 3 2
correct
chosen
spk/s
- Q4
δ and U
δ
U
trials
B
MOP2_5460 correlation : r=0.37 p<0.0001
post-Target mean firing rate (spk/s)
Q2
target 2
other
spk/s
trials (chronological order)
target ON
touch
(s)
target 1 4 2 4 2 1 2 1 3 4 2 3 2 1 4 1 3 2
correct
chosen
spk/s
Q2
δ and U
δ
U
trials

Figure 5

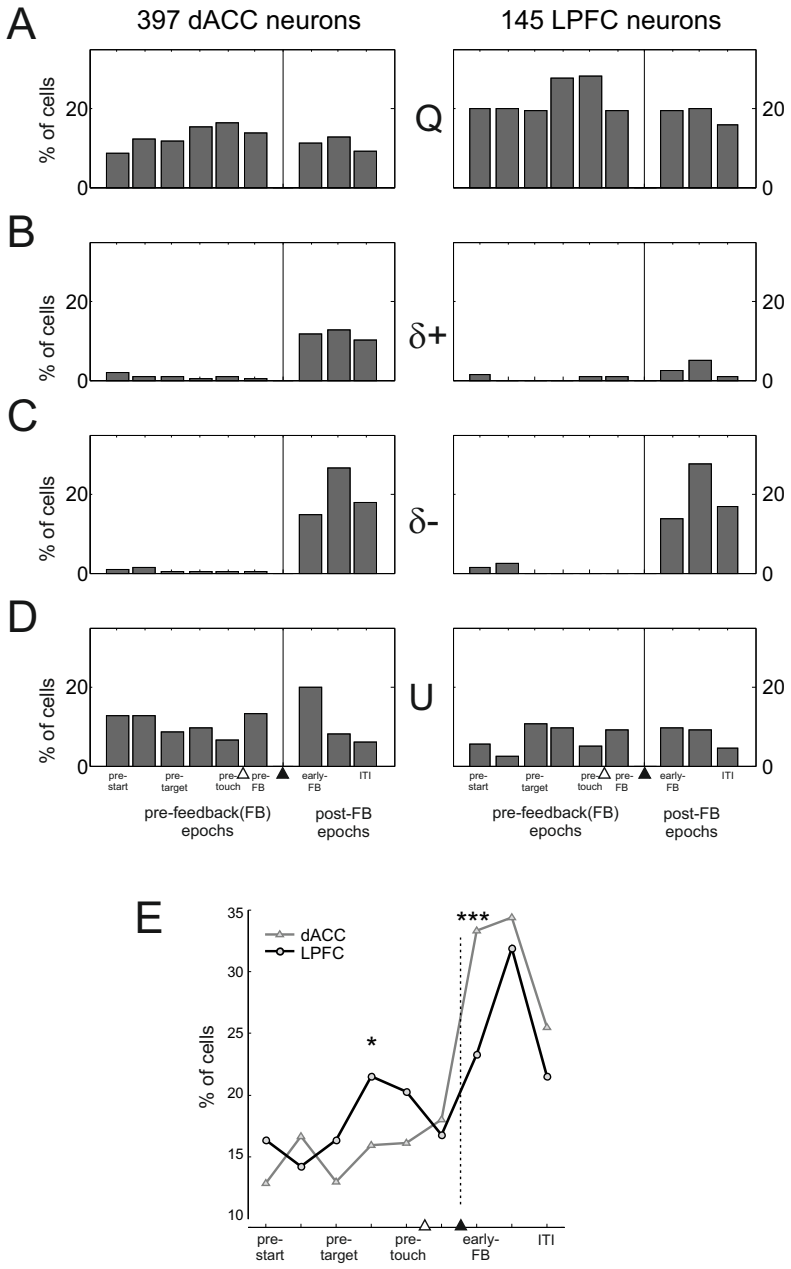

A
397 dACC neurons
145 LPFC neurons
% of cells
20
0
Q
B
δ+
C
δ-
D
U
pre-start
pre-target
pre-touch
pre-FB
early-FB
ITI
pre-feedback(FB) epochs
post-FB epochs
E
dACC
LPFC
35
30
25
20
15
10
*
***

Figure 6

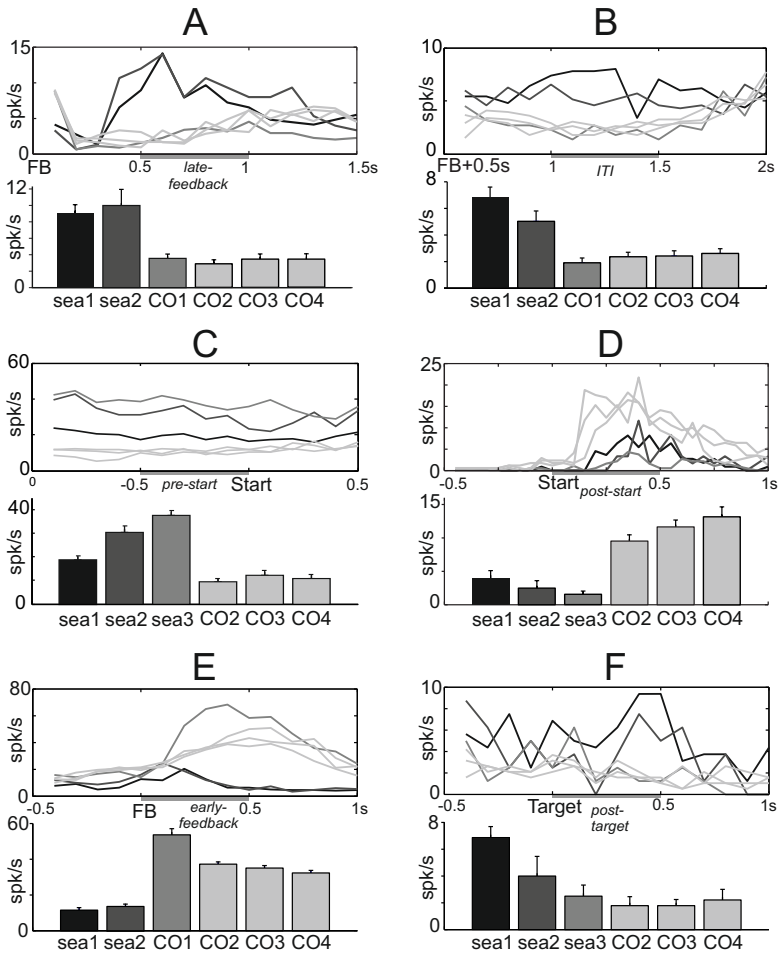

A
15
spk/s
0
FB
0.5
late-
feedback
1
1.5s
12
spk/s
0
sea1 sea2 CO1 CO2 CO3 CO4
B
10
spk/s
0
FB+0.5s
1
ITI
1.5
2s
8
spk/s
0
sea1 sea2 CO1 CO2 CO3 CO4
C
60
spk/s
0
0
-0.5
pre-start
Start
0.5
40
spk/s
0
sea1 sea2 sea3 CO2 CO3 CO4
D
25
spk/s
0
-0.5
Start
post-start
0.5
1s
15
spk/s
0
sea1 sea2 sea3 CO2 CO3 CO4
E
80
spk/s
0
-0.5
FB
early-
feedback
0.5
1s
60
spk/s
0
sea1 sea2 CO1 CO2 CO3 CO4
F
10
spk/s
0
-0.5
Target
post-
target
0.5
1s
8
spk/s
0
sea1 sea2 sea3 CO2 CO3 CO4

Figure 7

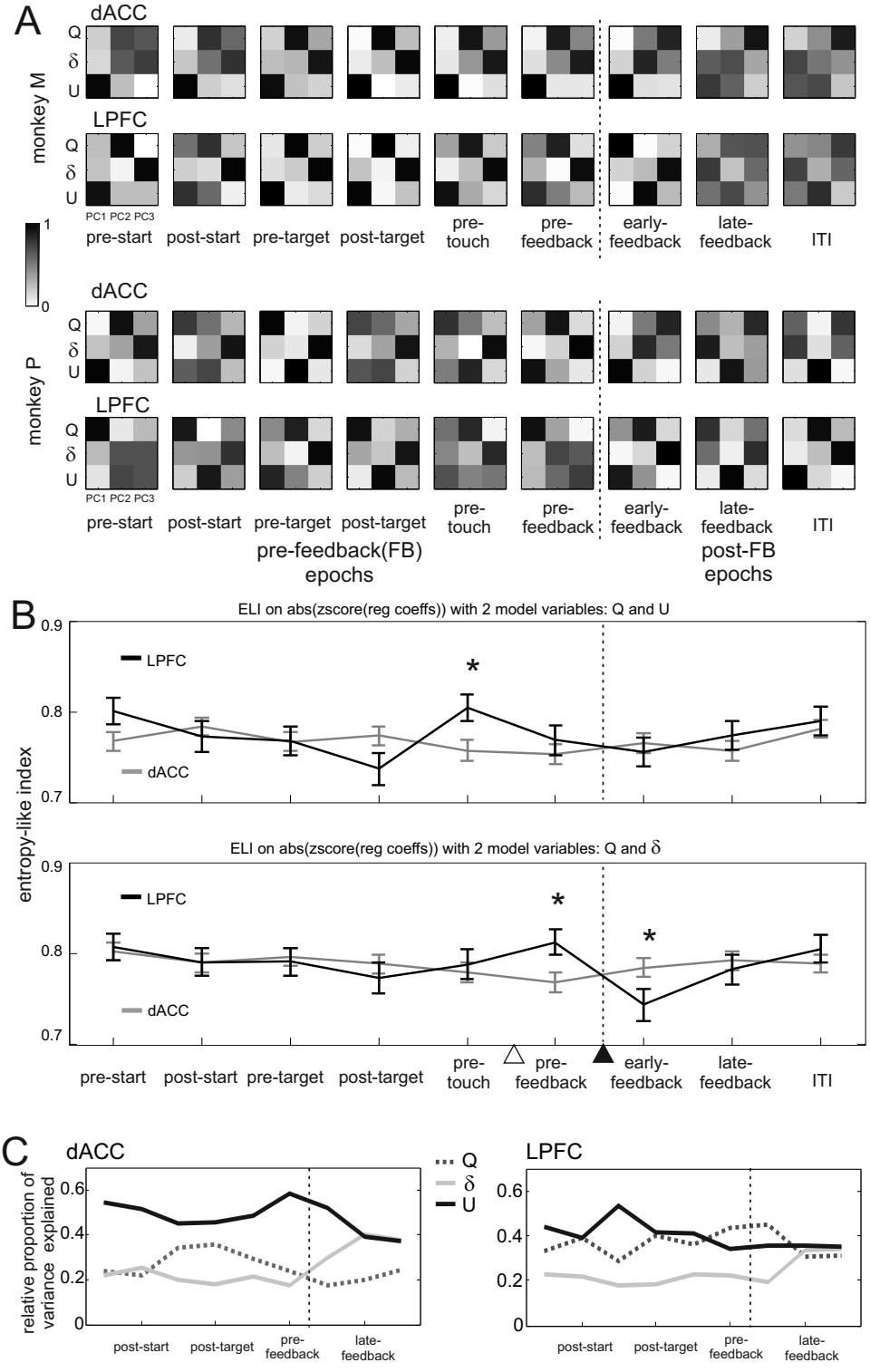

A
dACC
LPFC
monkey M
monkey P
Q
δ
U
PC1 PC2 PC3
1
0
pre-start
post-start
pre-target
post-target
pre-touch
pre-feedback
early-feedback
late-feedback
ITI
pre-feedback(FB) epochs
post-FB epochs
B
ELI on abs(zscore(reg coeffs)) with 2 model variables: Q and U
ELI on abs(zscore(reg coeffs)) with 2 model variables: Q and δ
LPFC
dACC
entropy-like index
0.9
0.8
0.7
*
C
dACC
LPFC
relative proportion of variance explained
0.6
0.4
0.2
0
post-start
post-target
pre-feedback
late-feedback

Figure 8

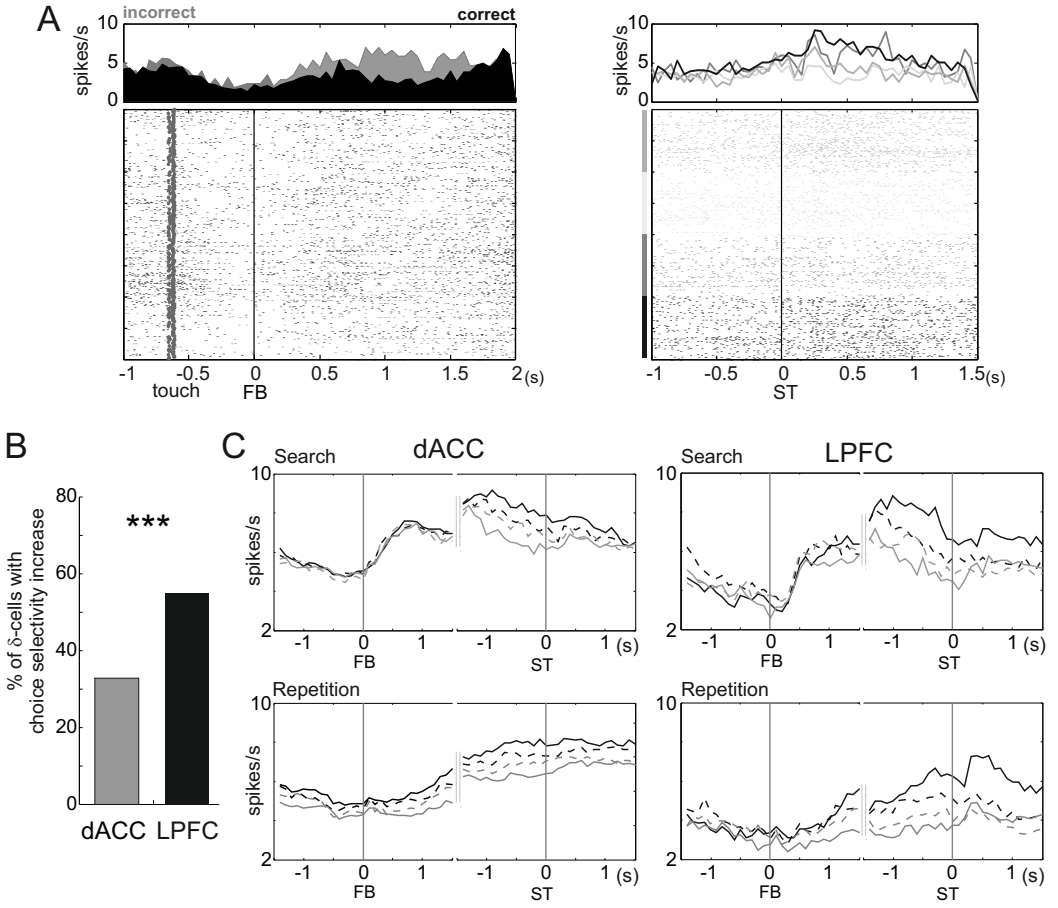

A
incorrect
correct
spikes/s
touch
FB
ST
B
% of δ-cells with choice selectivity increase
***
dACC
LPFC
C
Search
dACC
LPFC
Repetition
spikes/s
FB
ST

Figure 9

Behavioral regulation and the modulation of information coding in the lateral prefrontal and cingulate cortex

Mehdi Khamassi [1,2,3,4], René Quilodran [1,2,5], Pierre Enel [1,2], Peter F. Dominey [1,2], Emmanuel Procyk [1,2]

[1] Inserm, U846, Stem Cell and Brain Research Institute, 69500 Bron, France
[2] Université de Lyon, Lyon 1, UMR-S 846, 69003 Lyon, France
[3] Institut des Systèmes Intelligents et de Robotique, Université Pierre et Marie Curie-Paris 6, F-75252, Paris Cedex 05, France
[4] CNRS UMR 7222, F-75005, Paris Cedex 05, France
[5] Escuela de Medicina, Departamento de Pre-clínicas, Universidad de Valparaíso, Hontaneda 2653, Valparaíso, Chile

Supplementary table and figures

	dACC	LPFC
Multiple regression analysis		
Q cells	227 (39%)	126 (54%)
RPE cells	252 (44%)	69 (30%)
U cells	206 (36%)	48 (21%)
Cells w. multiple correlates	218 (38%)	75 (32%)
Cells w. single correlates	179 (31%)	70 (20%)
Cells without correlation	179 (31%)	87 (37.5%)
Cells w. correlates without other effect	*78 (14%)*	*20 (9 %)*
Mutual Info analysis		
Analysis on all cells		
Cells with M.I. < 0.1	409 (71%)	145 (62.5%)
Cells with M.I. > 0.1	167 (29%)	87 (37.5%)
Restrictive analysis (requiring a large number of samples)		
Excluded cells (not enough trials)	461 (80%)	159 (69%)
Included cells with M.I. < 0.1	111 (19%)	56 (24%)
Included cells with M.I. > 0.1	4 (1%)	17 (7%)
M.I. cells without other effect	*0 (0%)*	*0 (0%)*
SEA-REP activity variation analysis		
SEA<REP cells	96 (17%)	20 (9%)
SEA>REP cells	116 (20%)	39 (17%)
Non signif. variation cells	364 (63%)	173 (75%)
SEA<>REP cells without other effect	*22 (4%)*	*4 (2%)*
SEA-REP choice selectivity analysis		
SEA only selective cells	60 (10%)	12 (5%)
REP only selective cells	162 (28%)	83 (36%)
Both SEA and REP selective cells	64 (11%)	60 (26%)
Non selective cells	290 (50%)	77 (33%)
Choice selective cells without other effect	*27 (5%)*	*13 (6%)*
Non task-related cells	*61 (11%)*	*38 (16%)*
TOTAL number single units analysed	**576 (100%)**	**232 (100%)**

SUMMARY TABLE

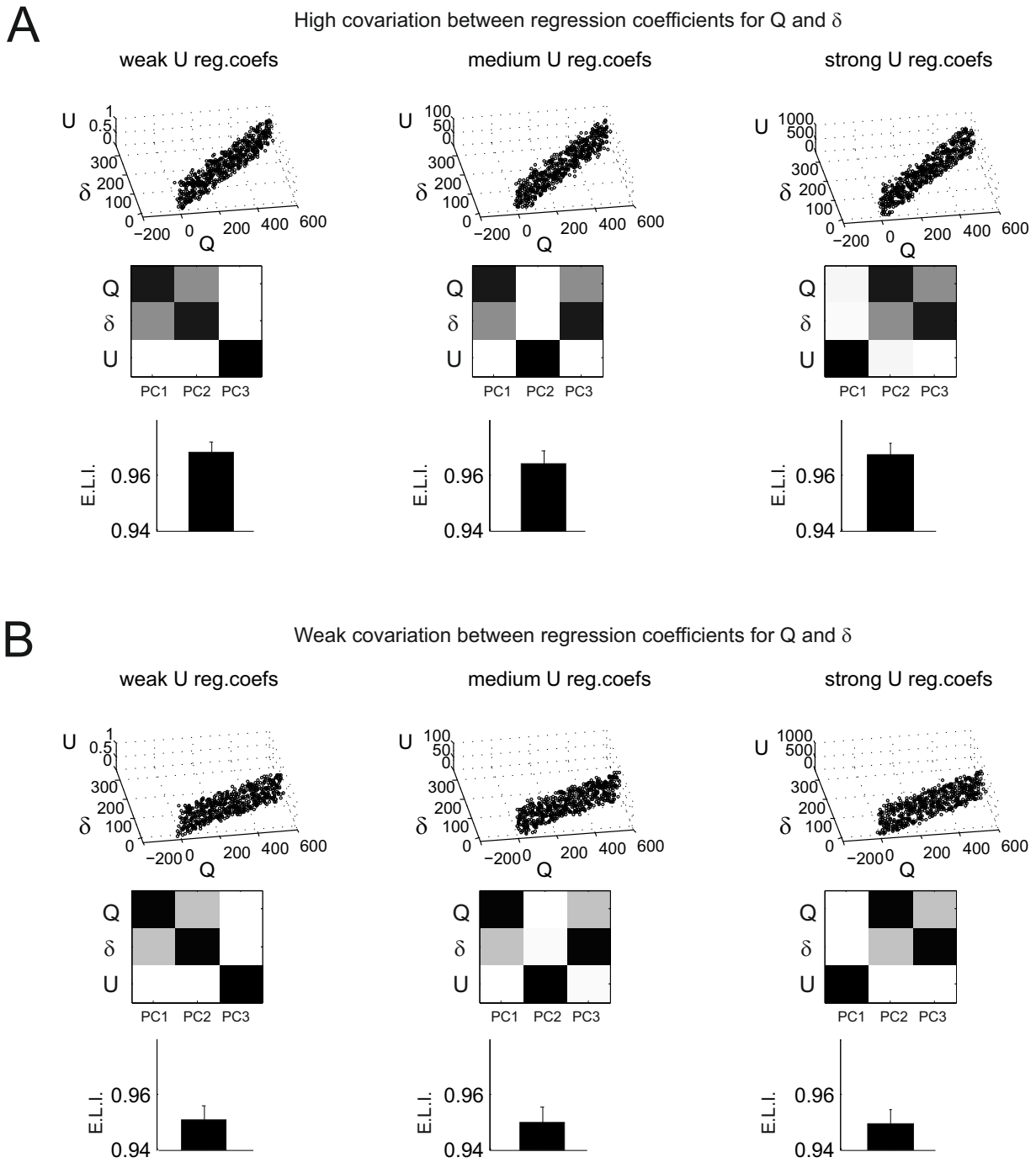


Figure S1. Simulations testing the effect of covarying variables. 6 ensembles of virtual data were created with covariations of coefficients of regressions (found with the multiple regression analysis cell x model variables) associated to Q and δ, and for which the coefficients associated to U are independant and represent a uniform noise (across the entire Z axis). The 6 data sets illustrate (from left to right, and from top to bottom) :

- case of strong covariation between coefficients for Q and δ, and weak reg coefficients associated to U (between 0 and 1)
- case of strong covariation between coefficients for Q and δ, and medium reg coefficients associated to U (between 0 et 100)
- case of strong covariation between coefficients for Q and δ, and strong reg coefficients associated to (between 0 et 1000)
- case of weak covariation between coefficients for Q and δ, and weak reg coefficients associated to U (between 0 and 1)
- case of weak covariation between coefficients for Q and δ, and medium reg coefficients associated to U (between 0 et 100)
- case of weak covariation between coefficients for Q and δ, and strong reg coefficients associated to (between 0 et 1000)

For each of the 6 cases 3 graphs are shown from top to bottom: - distribution of coefficients of regression for each of the 576 simulated cell data (one point per cell), - a matrix of the Principal Components (PC) for the three model variables (as in figure 8A), - the ELI (entropy-like index) measured on the absolute value of the Z-scores of the coefficients of regression associated to δ and Q.

These analyses show that the strength of correlation with model variables is reflected in the order of the principal components. They also show that strong covariation between regression coefficients for two different model variables results in principal components expressed as a function of both variables with nearly equal strength. These are the characteristics that are expected from the Principal Component Analysis applied to real neural data in dACC and LPFC.

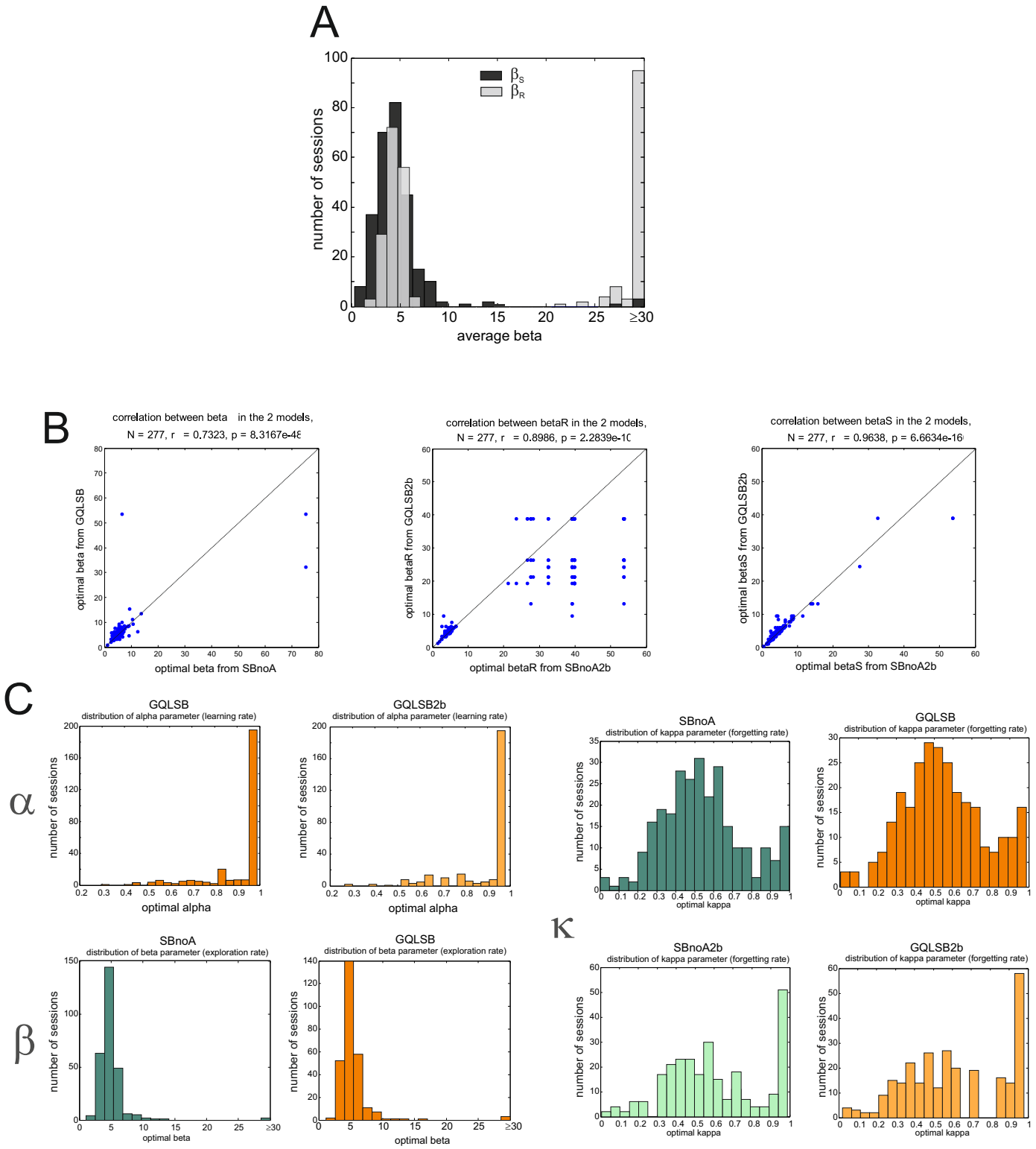


Figure S2. Distributions of Beta with model SBNoA and comparisons betwen GQLSB and SBnoA. A. Distribution of exploration meta-parameters obtained after optimization of the model on monkey's behavior using distinct degrees of freedom during the search period (β_s) and the repetition period (β_R). **B**. Comparisons of optimal βs obtained with SBnoA and GQLSB for one β versions, and 2 β versions. **C**. Distributions of meta-parameters (α, β, κ) over sesssions as obtained with the two models SBnoA and GQLSB, with one or 2 β as indicated on the figures. Green is for SBnoA, orange for GQLSB. Overall the figures shows the high similarity between the two models in their capacity to describe behaviour.

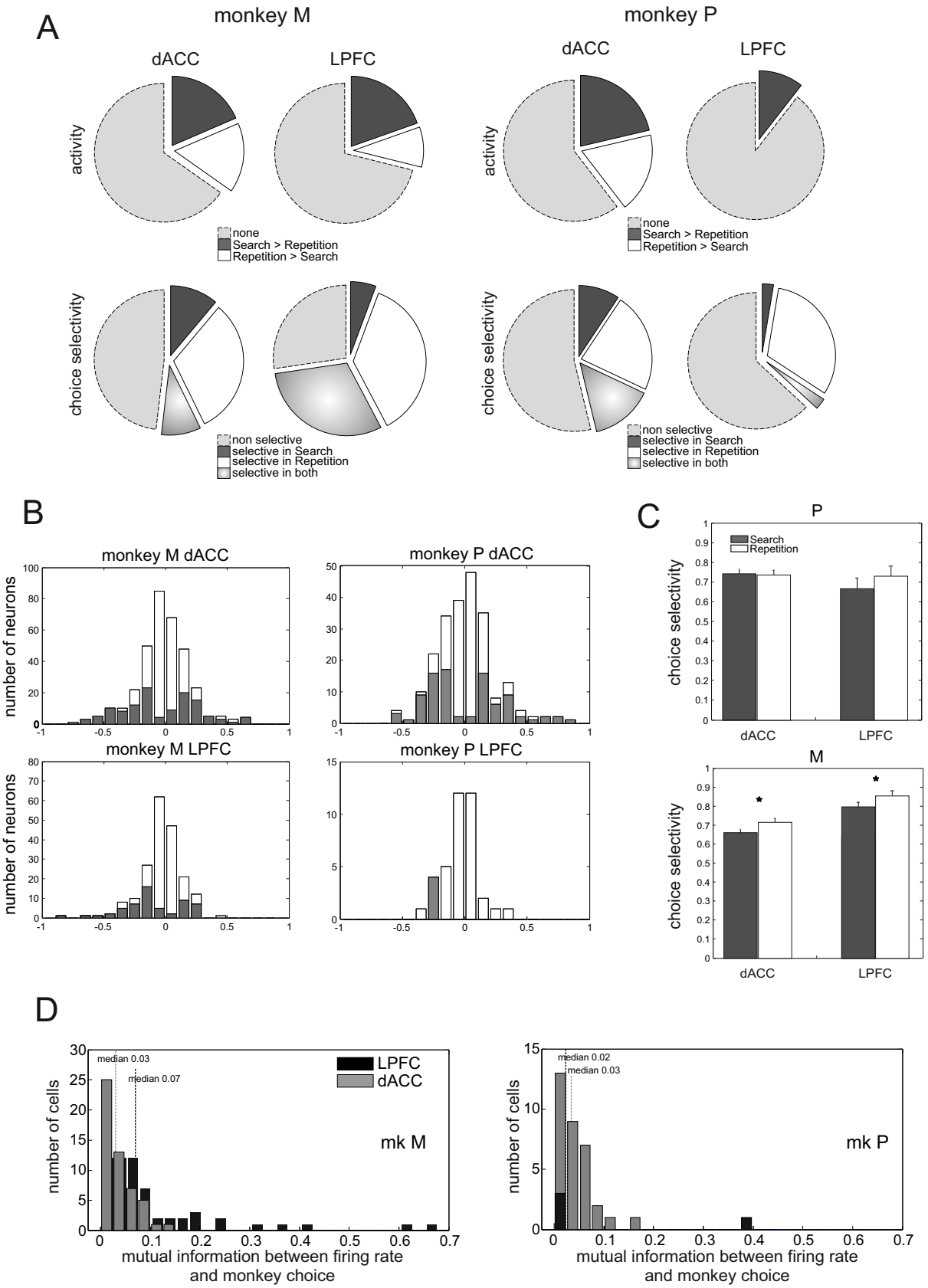


Figure S3. Variations of early-delay activity and choice selectivity - data for each monkey (M and P). (A-top) Proportions of dACC and LPFC cells with a higher activity during search (Sea) or repetition (Rep). **(A-bottom)** Proportions of dACC and LPFC cells with a higher choice selectivity during Sea or Rep. **(B)** Number of cells with significant changes (in grey) in average unit activity between search (Sea) and repetition (Rep). **(C)** Increase of choice selectivity from search to repetition in the two structures. Stars indicate statistically significant comparisons *: p<0.05, **: p<0.01. **(D)** Mutual information between the early-delay average firing rate and the animal's choice. Dashed grey and black lines represent the medians for dACC and LPFC respectively.

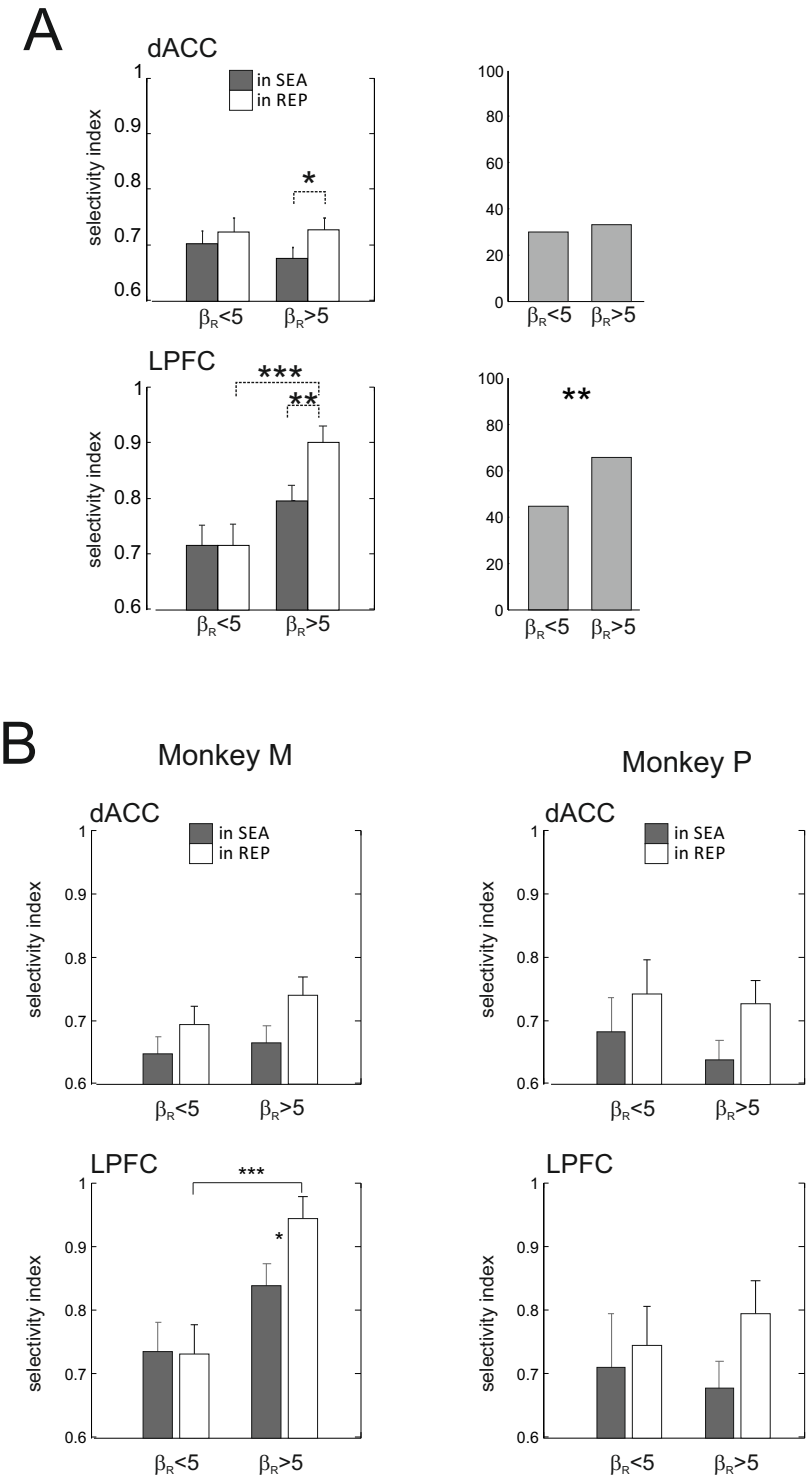


Figure S4. A. Choice selectivity and exploration level. Data computed using the SBNoA2β model (Left), and proportion of dACC and LPFC early-delay choice selective neurons during repetition periods of sessions where β_R was small (<5) or large (>5) (obtained with model SBNoA2β- Right). **B. Choice selectivity depending on exploration level using model GQLSB 2 Beta for each monkey (M and P).** The average choice selectivity index is presented for units recorded in dACC (top) and LPFC (bottom), in sessions grouped according to the fitted model's exploration parameters for search (β_S) and repetition (β_R). The average population index is measured for search (grey bars) and repetition (white bars) trials in the early-delay epoch, separately for sessions where β_S was inferior or superior to 5, and for sessions where β_R was inferior or superior to 5. Stars indicate statistically significant comparisons. *: $p<0.05$. When separating the data for the two monkeys, no significant effect was found in dACC for neither monkeys (Kruskal-Wallis test with Bonferroni correction, $p > 0.05$), a significant effect of β_R was found in Monkey M LPFC (Kruskal-Wallis test with Bonferroni correction, $p < 0.05$), and a tendency, although non-significant, was found in Monkey P LPFC.

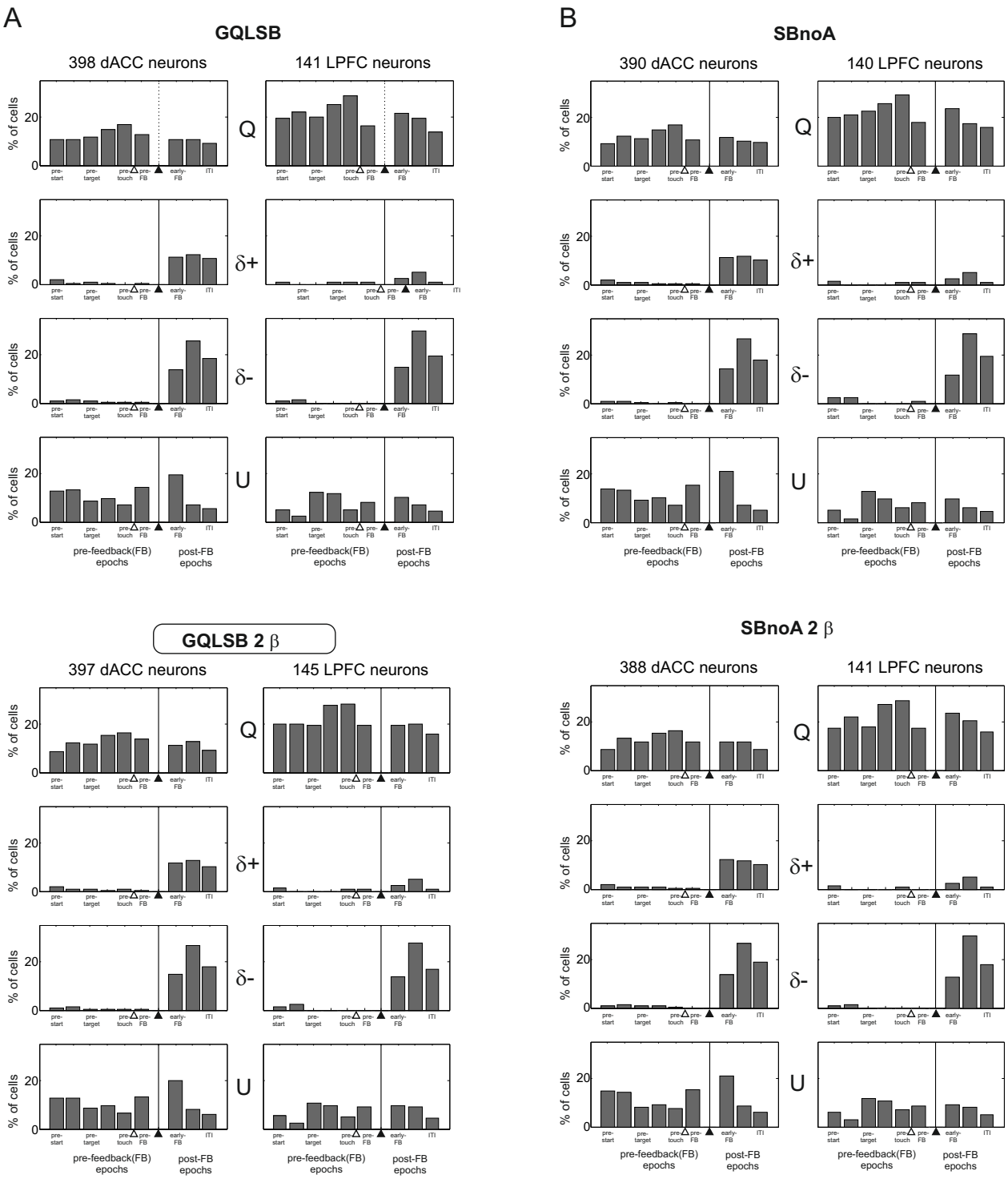


Figure S5. Proportions of dACC and LPFC cells with activity correlated with one of the model variables (Q, δ, and U) using 4 different models. The GQLSB model (A), and the SBNoA model (B) with 1 or 2 β parameter. (top and bottom). The GQLSB 2β is the model used for further analyses and presented in main figure 6.

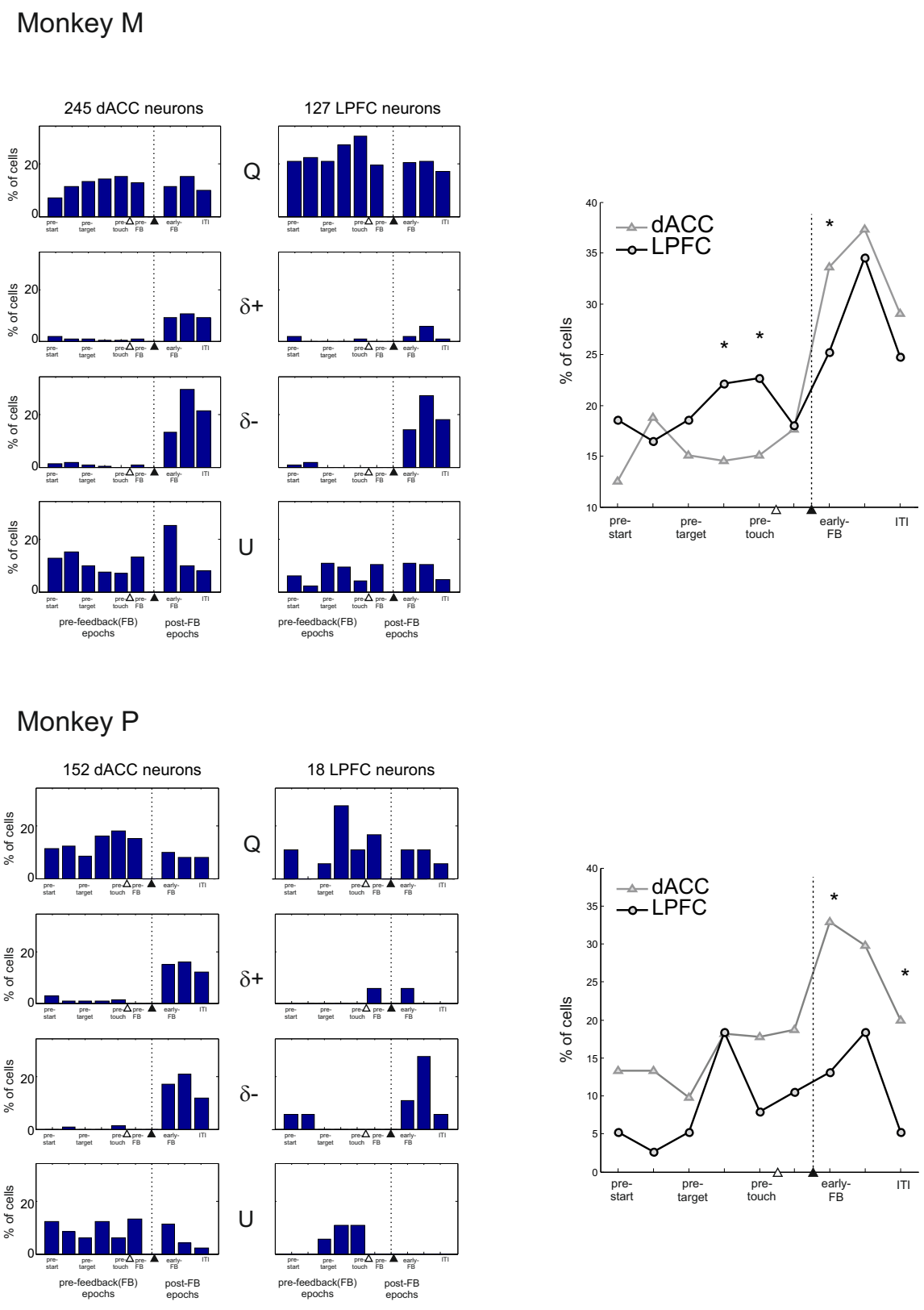


Figure S6. Proportions of dACC and LPFC cells with activity correlated with one of the model variables (Q, δ, and U) using the GQLSB 2 β model for each monkey (**Left**). On the **Right**, Proportion of cells, for each epoch, showing a significant correlation with at least one model variable. See figure 6 for average data and figure S5 for comparisons with other models.

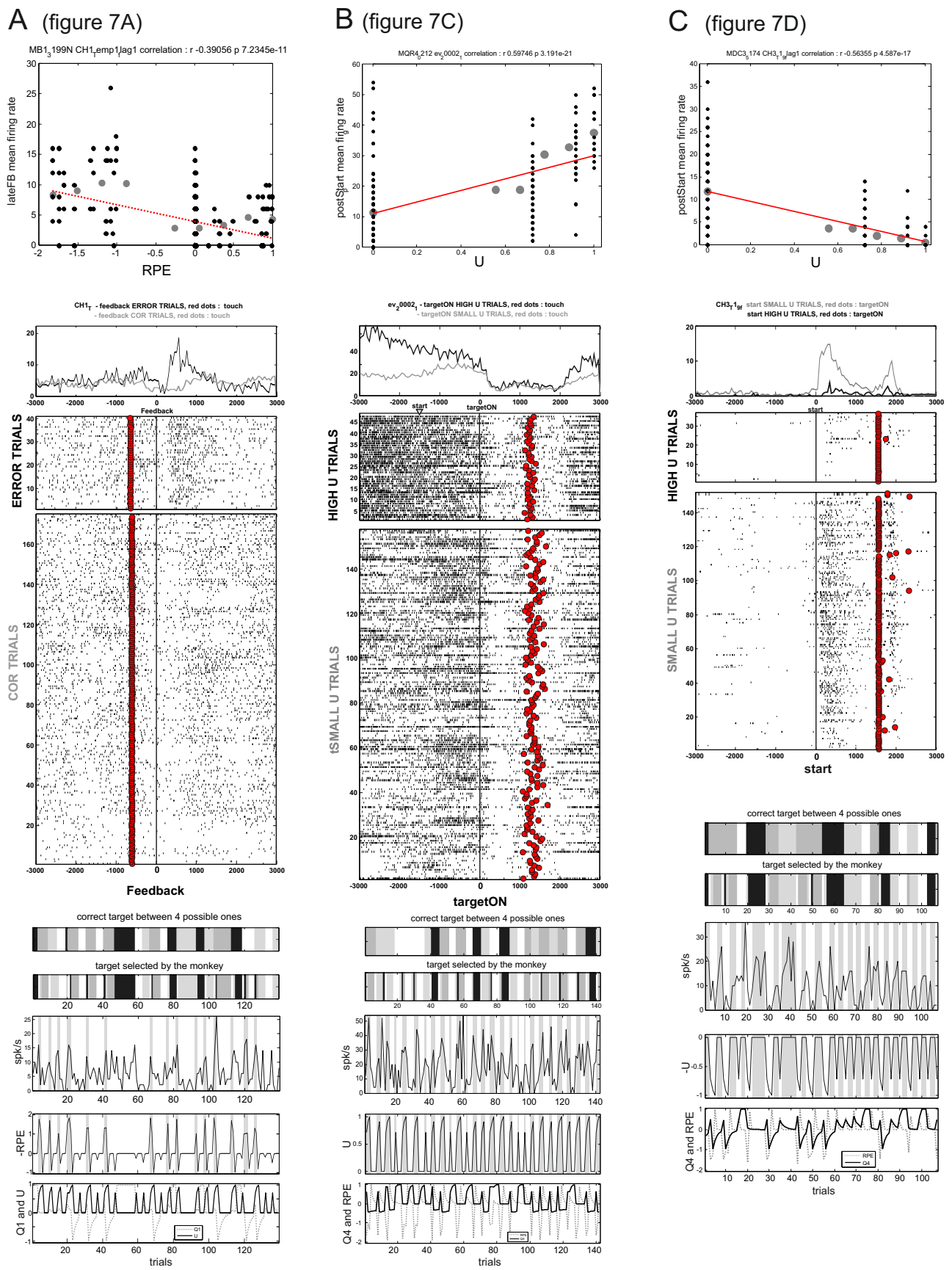


Figure S7. Three examples of unit activity from figures 7A (A), 7C (B) and 7D (C) correlated with some of the model's variables. (A) example of dACC activity negatively correlated with RPE (δ-). **(B)** example of LPFC activity correlated with U. **(C)** example of dACC activity negatively correlated with U. (**Top**) plot of single trial activity (black dots) measured in the late feedback (A) and post-Sart (B, C) epochs against RPE and U values respectively. Large grey dots represent the average for one decile of the value distribution and are just used for illustration. The red line represents the linear regression computed from single trial data. (**Middle**) peri-stimulus histograms aligned on feedback (A), Target Onset (B), and Start (C) and the corresponding raster plots for trial types indicated on the figures. (**Bottom**) trial by trial evolution of the average activity measured in the relevant epoch during successive trials in the session. The upper grey barcode represents the correct target to be chosen (4 greys for 4 target positions). The second barcode represents the target chosen by the animal in each trial. Below, the graphs represent the average activity for each trial and the trial by trial evolution of key model variables.

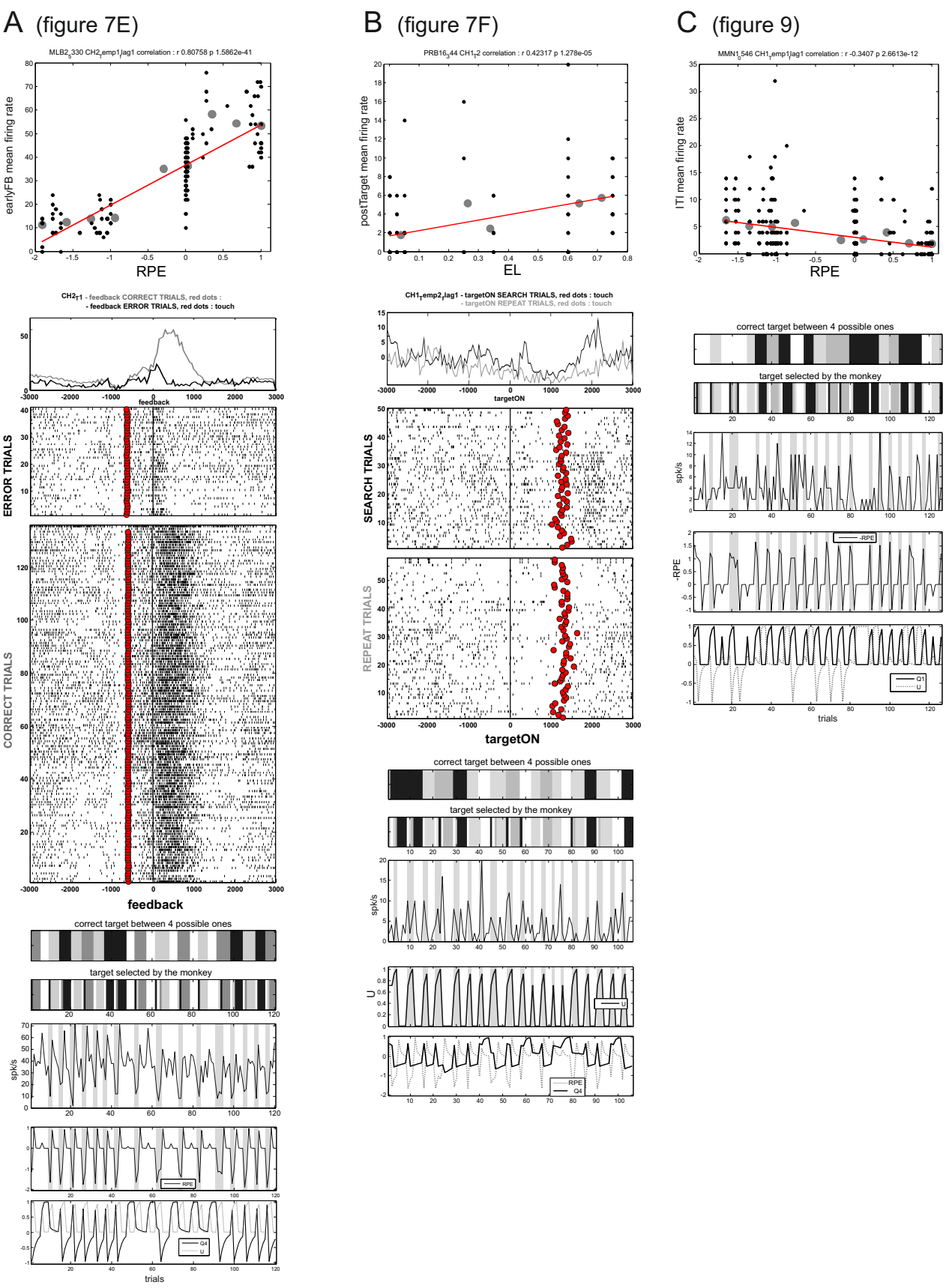


Figure S8. The two exemples from figures 7E (A) and 7F (B) correlated with some of the model's variables. **(A)** example of dACC activity positively correlated with RPE (δ+). **(B)** example of activity discriminating search and repetition but with a different profile than U; profile labelled EL for Error Likelihood. (**Top**) plot of single trial activity (black dots) measured in the early feedback (A) and post-target (B) epochs against RPE and EL values respectively. Large grey dots represent the average for one decile of the value distribution and are just used for illustration. The red line represents the linear regression computed from single trial data. (**Bottom**) peri-stimulus histograms aligned on feedback (A) and Target Onset (B) and the corresponding raster plots for trial types indicated on the figures. Other conventions as in Fig S7.

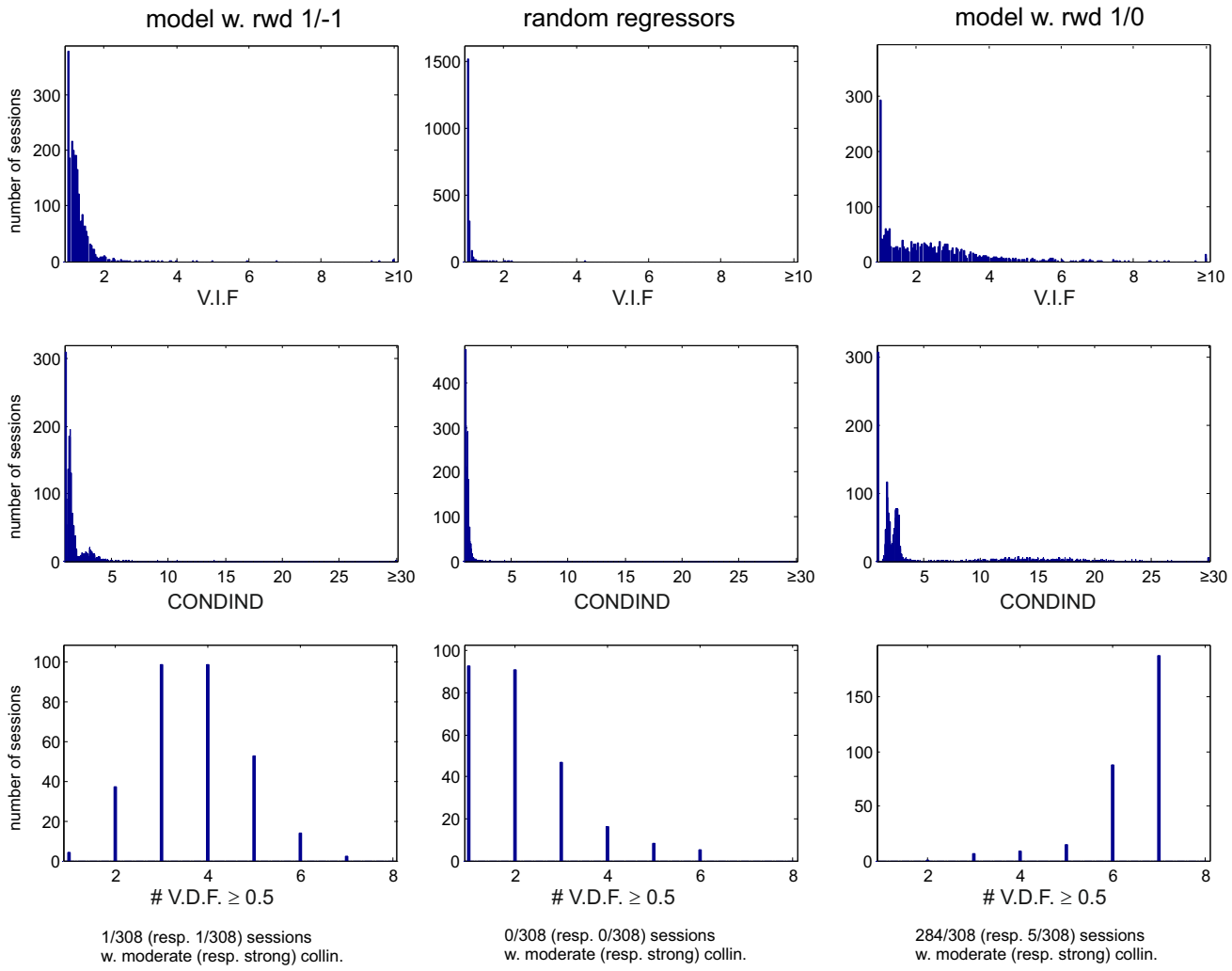


Figure S9. Analyses of colinearity. Evaluation of the degree of collinearity between regressors used in the multiple regression analysis of single-unit activities as a function of model variables. (**Left**) Model GQLSB2β with the reward function used throughout the paper (1 in case of success, -1 in case of failure); (**Middle**) Control model with randomly generated regressors; (**Right**) Model GQLSB2β with a different reward function (1 in case of success, 0 in case of failure). For each recording session (308 in total) and for each regressors (7 in total), the figure shows the degree of collinearity measured when expressing the regressor as a function of the 6 other regressors for that session.
The histograms on **top** show the variation inflation factors (**VIF**) computed with the coefficient of determination obtained when each regressor was expressed as a function of the other regressors. The **middle** figure shows the condition indexes (**CONDIND**) obtained in the same analysis. The bottom figure shows the number of variance decomposition factors (**VDF**) superior or equal to 0.5 obtained for each recording session.

The figure shows that the GQLSB2β model used throughout the paper (Left) displayed a strong collinearity between regressors only for 1/308 session (condind>=30 and more than two VDPs > 0.5) and a moderate collinearity only for 1/308 session (condind>=10 and more than two VDPs > 0.5). All other sessions showed a weak collinearity between regressors. In contrast, when the same model is used with a reward function equal to 1 for correct trials and 0 for error trials, collinearity is strong for 5/308 sessions and moderate for 284/308 sessions. As a control, a model with randomly generated regressors shows weak collinearity in 100% simulated sessions.

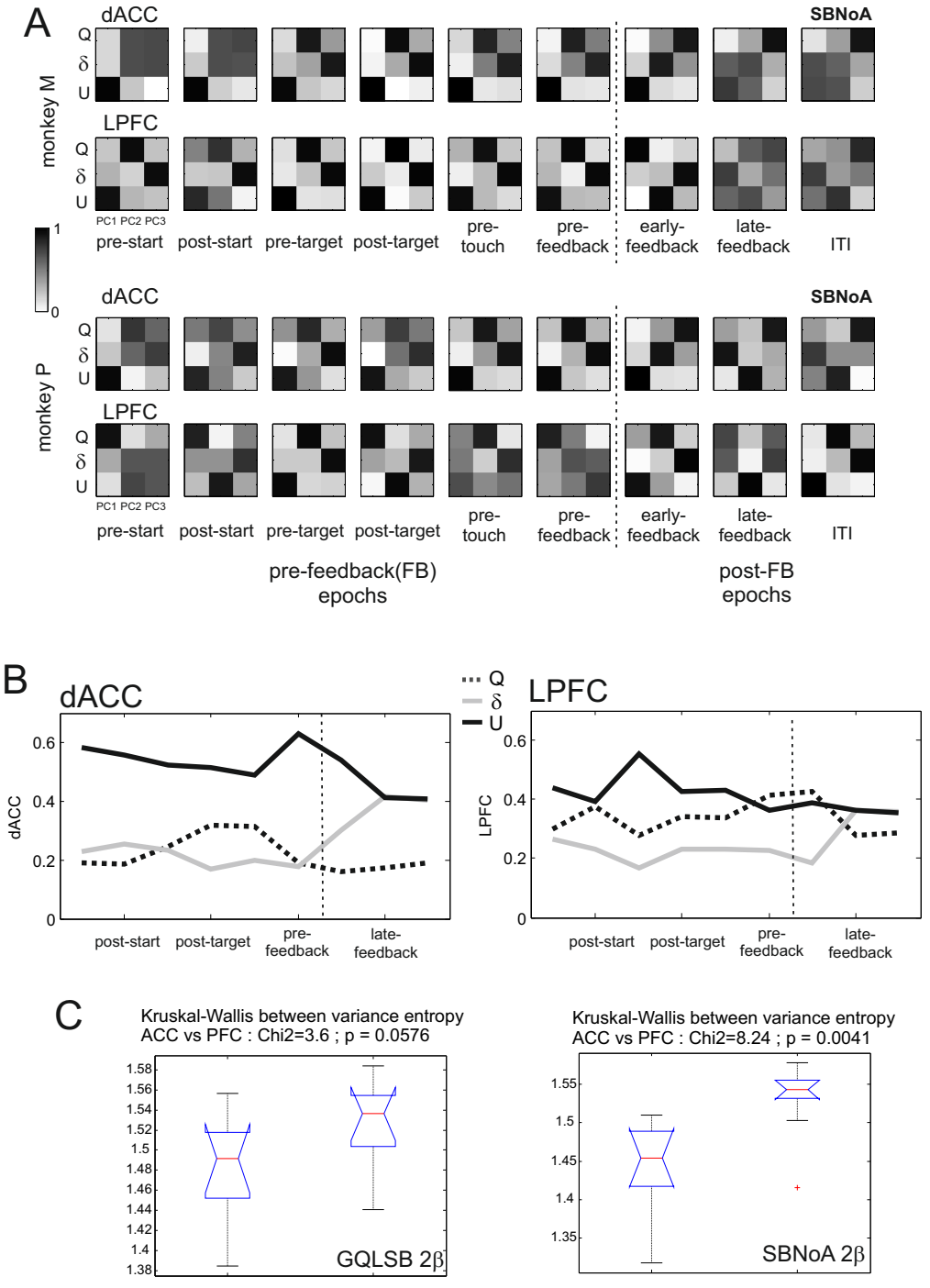


Figure S10. Multiplexing of information and variations during trials in dACC and LPFC - Data given for model SBNoA2β. (**A**) A principal component analysis was performed on the regression coefficients found for each neuron and for each model variable (Q: the action value of the animal's preferred target, δ, and U). The absolute value of the eigen values for each principal component computed during the early-feedback epoch are shown in each matrix for one trial epoch. **(B) Top**. Proportion of total variance explained by each model variable over the 3 PCs for dACC and LPFC data along trial epochs. **Bottom**. Comparison between models GQLSB2β and SBnoA2β of an entropy-like index computed on the set of % variance explained by each model variable in each trial epoch (data from part A). A kruskal-Wallis test indicated a higher entropy in LPFC than in dACC (marginal significance for model GQLSB2β; strong significance for model SBnoA2β). See main text for details.

3.2 Dopamine activity during decision-making in rats

3.2.1 Bellot, Sigaud, Roesch, Schoenbaum, Girard, Khamassi (in prep)

What do VTA dopamine neurons encode: value, RPE or other behaviour correlates?

Jean Bellot, Olivier Sigaud, Matthew Roesch, Geoffrey Schoenbaum, Benoît Girard, Mehdi Khamassi[a]

[a] *Université Pierre et Marie Curie,*
Institut des Systèmes Intelligents et de Robotique - CNRS UMR 7222,
Pyramide Tour 55 - Boîte Courrier 173,
4 Place Jussieu, 75252 Paris CEDEX 05, France
name.lastname@isir.upmc.fr

Abstract

Traditionally, dopamine neurons are hypothesized to encode a reward prediction error which is used in temporal difference learning algorithms. This hypothesis is based on numerous studies that qualitatively analyzed the activity of dopamine neurons during learning. However, the exact nature of such signal is still unclear, notably when the task involves multiple choice. In order to further investigate the parallel between these two information, we simulated standard temporal difference algorithms in a multi choice task, which has been used for electrophysiological recordings of dopamine neurons, in order to investigate their ability to reproduce the pattern of previously recorded dopamine signal. We used a quantitative method that enables direct comparison between simulated reward prediction error signal and dopamine activity. Our results indicate that the dopaminergic signal could not be accurately reproduced by a pure reward prediction error signal and seems to embody value function information. Furthermore we show that the information carried out by dopamine neurons seems to be at least partly dissociated from behavioral adaptation.

1. Introduction

During the 90's, the work of Schultz and colleagues [18, 27, 30, 36] has led to major progress in understanding the neural mechanisms underlying the influence of feedback on learning. In these studies, the activity of dopaminergic (DA) neurons exhibited four key properties of the reward prediction error (RPE) signal used in so-called Temporal Difference (TD) machine learning algorithms [10, 37, 39]: (1) they responded to unexpected rewards; (2) they responded to reward predicting cues (conditioned stimuli, CS); (3) they did not respond to expected rewards; (4) they showed a decrease in activity in response to omission of an expected reward. This RPE signal acts as a teaching signal, allowing TD learning algorithms to learn to predict future rewards based on current state and action. Using this signal, algorithms update their prediction of reward and eventually learn to predict the amount of reward they should get in the future. Considering the strong connectivity between the DA system and the basal ganglia known for its action selection properties [30], DA has thus been thought to be the neural signal that help us to adapt our behavior based on trials and errors.

This hypothesis has been confirmed and extended by numerous studies showing the relevance of TD learning algorithms to the mechanisms of action selection and behavioral adaptation involving DA neurons and the basal ganglia [2, 13, 33, 40]. However, the precise information encoded by DA signals remains unclear. One reason for this is that DA activity has been primarily recorded during tasks where the animal is passive, thus the results cannot reveal the link between this signal and the choice of an action. This is important because different TD learning algorithms treat the importance of behavior or actions differently. More recent electrophysiological studies have addressed this issue, measuring DA activity during multi-choice tasks. However, these studies arrived at divergent conclusions concerning which algorithm best explains the influence of action on DA activity. One approach found that DA activity reflected future choices [31, 32] consistent with predictions of Sarsa algorithms, while another approach found that DA activity reflected the best available option irrespective of future choices [6, 8, 34], consistent with predictions of Q-learning. Addition to the confusion, the known anatomy of the basal ganglia suggests an architecture closer to the Actor-Critic [20].

In this study, we aimed to resolve these issues by analyzing more precisely and more quantitatively the information encoded by DA neurons, using a dataset from one of these above studies. In this study by Roesch et al. 2007 [34], DA neurons were recorded in rats cued to choose between two actions leading to differently delayed and sized rewards (Fig. 1). During some trials, termed free-choice, two different rewards were accessible, and the rats had to learn to choose the action leading to the most attractive reward. After a few trials, rats were able to choose the immediate or the big reward more often than the delayed or small reward. Prior analyses of the main characteristics of DA neurons' activity averaged over post-learning trials [34] suggested that the DA signal pattern looked similar to the RPE computed by the Q-learning algorithm: the amplitude of response to the cue in free choice trials was the same no matter the value of the action actually performed by the animal, and this amplitude was not different from the maximal amplitude observed during forced choice trials.

*corresponding author, email: mehdi.khamassi@isir.upmc.fr phone: (+33).1.4427.8853, fax: (+33).1.4427.5145

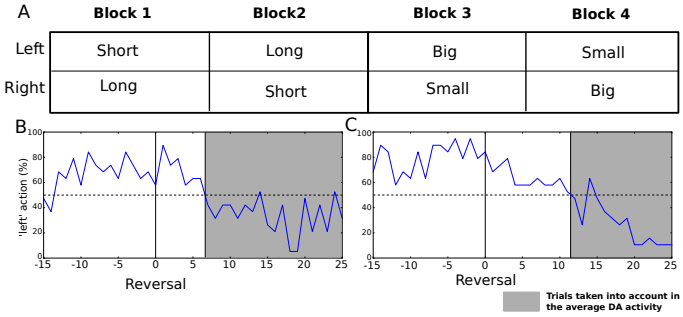


Figure 1: Description of the task of Roesch et al. 2007. A. Each session is composed of 4 different blocks. Each block has a different contingency and block changes are unsignalled. The first two blocks are the delay blocks. In the first one, the short reward is delivered in the left well and the long reward is delivered in the right well. The second block has the opposite contingency. Blocks 3 and 4 are the size blocks. In block 3, the big reward is delivered in the left well and the small reward in the right well. Block 4 has the opposite contingency. B-C. Animal's behavior recorded during the size and delay reversal respectively. In grey are represented the trials from which DA activity has been recorded.

However, a closer examination of the Figures of [34] reveals other characteristics that are inconsistent with the RPE hypothesis: the post-learning DA response to expected reward is higher than response to the cue (unlike an RPE signal that would have already converged), and there is no dip in DA response to smaller than expected rewards in some trial blocks. While the main characteristics of DA response seem consistent with an RPE signal, the latter characteristics appear to better correspond to a value function. Therefore, in this study, we performed systematic simulations of the main candidate reinforcement learning algorithms and extracted both RPE and value information in order to test whether DA activity reflects a pure RPE signal, a pure value signal or a mixture of the two. Interestingly models with only a pure RPE signal failed to reproduce the observed DA activity patterns, showing the limit of the link between DA activity and the RPE signal calculated by TD learning algorithms.

We also tested the importance of behavior in explaining the firing of the DA neurons. We found that constraining the algorithms to fit both behavior and DA activity degraded the fit between the models and the neural activity patterns. In contrast, releasing the constraint to fit behavior enabled a mixture of value and RPE calculated by the Actor-Critic model to fit DA activity well. Overall these results suggest that a more complex interaction between learning to predict reward and behavioral adaptation, such as that proposed in dual learning system models [7, 22], is required to reproduce the DA activity observed in Roesch et al. [34] work.

2. Material and methods

2.1. Experimental procedure

In this task, rats perform blocks of trials where they must learn to choose the best option between two wells delivering various rewards (see Figure 1). In blocks 1 and 2 called delay blocks, one well is associated with an immediate reward (*short* option), the other one with a delayed reward (*long* option). In order to prevent the animal from giving up if it experiences a sudden high delay, the duration of the *long* option is progressively increased: 1 sec at the first trial where the animal selects the *long* option, 2 sec at the second trial, until 7 sec maximum. In contrast, if it chooses more than 8 times over the last 10 trials the path to the *short* option, then the delay for the *long* reward is shortened. In blocks 3 and 4 called size blocks, one well is associated with a large reward (*big* option), the other one with a small reward (*small* option; see Figure 1). Blocks are organized so that the best option is alternatively left or right: e.g. left = *short* during block 1, left = *long* during block 2, left = *big* during block 3, left = *small* during block 4. Block changes are not signalled, forcing rats to learn to switch their preferred well from their own errors. Thus, in each block, rats must choose between the left and the right well, and learn by trial and error which well conveys the best benefit/cost ratio (i.e. big reward in the size case and short term reward in the delay case).

One odor among three is presented at each trial to help the rat making its choice. This odor is the conditioned stimulus (CS) with which the rewarded well is associated. Odor 1 always indicates that the left well contains a reward (short, long, big, small depending on the current block) while the right well is empty. Odor 2 always indicates that the right well contains a reward (short, long, big, small depending on the current block) while the left well is empty. Thus trials where odor 1 or odor 2 are presented are called *forced choice* trials because the animal can only get rewarded with a single option. Odor 3 indicates that reward can be found on both sides, the quality of the reward depending on the current block. Thus trials where odor 3 is presented are called *free choice* trials.

While the rats experienced these various blocks, the experimenters recorded the behavior and the activity of putative DA neurons in the ventral tegmental area (VTA). Two additional DA neurons were recorded from the substantia nigra pars compacta (SNc). Neurons were identified as dopaminergic using both the waveform criteria and the impact of an injection of the DA agonist apomorphine on their activity (more details can be found in the original experiment [34]).

2.2. Modelling the experimental task

We modeled the experiments of Roesch et al 2007 [34] with a Markov Decision Process (MDP) (see Figure 2 A). Each state represents a salient event in the original task that triggers a phasic DA response: the beginning of the trial, the nosepoke, the perception of the odor and the delivery or omission of the reward (see Figure 2 B). Thus there is a correspondence between the states of the MDP and the events experienced by rats. The transition between the nosepoke state and the *odor* state is not action-dependent but is generated by the simulation in order to present each odor the same amount of time, as in the original experiment.

The reward states labeled "R{L for left or R for right}{n of odor}{$i \in [1:14]$ represent the delay}", e.g. RL31 on Figure 2A, represent a succession of states modeling the reward (see Figure 2 C and D). The model accounts for all the rewarding schemes, i.e. the *delay* case and the *size* case. In order to switch

to another block, the simulation manipulates the delay of the reward (in the *delay* case; see Figure 2 C) or adds a new reward (in the *size* case; see Figure 2 D). The delay is modeled as a succession of states without reward, and one transition between two states corresponds to 0.5s in the real task.

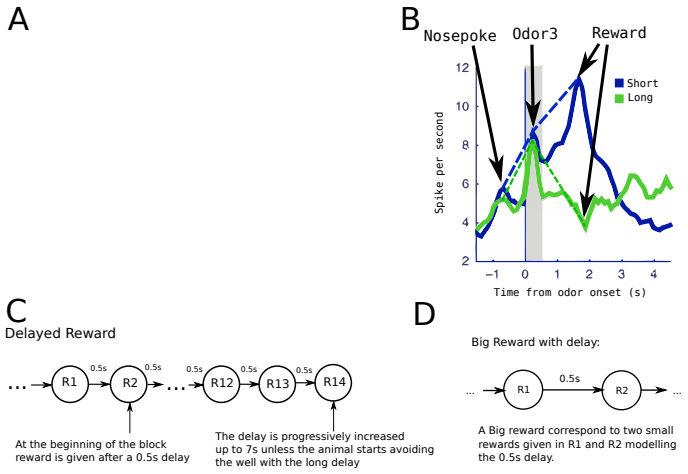


Figure 2: Modeling the state of the tasks used in [34]. A. Markov Decision Process used to model the task; RL3, Reward Left following odor 3; RR3, Reward Right following odor 3. The other states represent the delay. B. State decomposition illustrated on the DA activity reported by Roesch. We extracted the DA activity from the original recording during the three salient events of a trial : the time of the 'nosepoke', the perception of the 'odor' and the rewards state RL or Rn of the odor. C. Model of the long reward. D. Model of the big reward as two small rewards.

Furthermore, the *big* reward is modeled as two *small* rewards delivered in two consecutive states (see 2 D). It simulates the slight delay of 0.5s between the two *small* rewards in the original experiment.

Since odor 1 and odor 2 are forced choices indicating a reward on the right and left respectively, no reward is given on RL1 and RR2. Moreover, as in the original study, the rewarding scheme is the same in RR1 and RR3 and in RL2 and RL3.

The value of the reward is set to 5 in our simulation to model the 0.05-ml bolus of 10% sucrose solution given to rats.

2.3. Studied algorithms

The general RL used in the model is shown in Algorithm 1. We compared three algorithms: Q-learning, Sarsa and Actor-Critic. Q-learning and Sarsa are based on the same principles. They update for each state action pair (s, a) a Q-table that stores the expected utility for performing that action a in state s.

This Q-table is called a *critic*: it tells you how good it is to choose a particular action in any state. This information is sufficient to decide what to do in any situation, so an agent does not need a further structure to determine its policy.

In contrast, the Actor-Critic architecture also contains a *critic*, but additionally it contains a different structure called the actor which represents the policy of the agent. In the version studied here, the critic contains less information than in Q-learning and Sarsa. Instead of storing in a Q-table the expected utility for performing all actions in state s, it only stores in a V vector the expected utility of each state s.

The actor is represented as a table $\mathcal{P}$ which associates to any (s, a) pair a value corresponding to the probability of performing action a in state s, i.e. $\mathcal{P}(s, a) \propto P(a|s)$. The action actually chosen is determined through a *softMax* function.

The critic structures, Q and V, are updated from the *TD error* δ using $\forall f \in \{Q, V\} : f_{t+1} = f_t + \alpha\delta_t$. But the computation of the TD error differs depending on the algorithm:

- Q-learning: $\delta_t = r_{t+1} + \gamma \max_a(Q(s_{t+1}, a)) - Q(s_t, a_t)$
- Sarsa: $\delta_t = r_{t+1} + \gamma(Q(s_{t+1}, a_{t+1})) - Q(s_t, a_t)$
- Actor-Critic: $\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$

In Sarsa, the critic is updated given the value of the action that will be performed in the next state, whereas in Q-learning it is updated assuming the agent will take the best action in the next state, without requiring that this assumption is verified in practice. This distinction is critical in the analysis of the dopaminergic signal performed in [31, 32, 34].

In the Actor-Critic architecture, the actor $\mathcal{P}$ must also be updated using $\mathcal{P}_{t+1} = \mathcal{P}_t + \alpha\prime\delta_t$, where the learning rate $\alpha\prime$ can be different from the one used to update the critic. For instance, if $\alpha\prime$ is smaller than α, this may induce a slower convergence of behavior with respect to the *TD* error, which is assumed to correspond to dopaminergic signal in this paper. More generally, having a different structure for the actor and for the critic, the Actor-Critic architecture makes it easier to represent a behavior that is not under strict control of the critic.

In order to choose the performed action, the same *softMax* policy π is used whatever the algorithm to choose an action:

$$\pi(a|s_t) = \frac{\exp(\beta Q(s_t, a) \; or \; \mathcal{P}(s_t, a))}{\sum_b \exp(\beta Q(s_t, b) \; or \; \mathcal{P}(s_t, b))}.$$

Algorithm 1 Learning

Require: initial state: s_0, block: mdp

$s_t \leftarrow s_0$
$a_t \leftarrow softMax(s_t)$
for $i = 0$ to max_iter **do**
  while $s_t nonterminal$ **do**
    $s_{t+1} \leftarrow Transition(s_t, a_t)$
    $a_{t+1} \leftarrow softMax(s_{t+1})$
    update $Q(s_t, a_t)$ or $[V(s_t)$ *and* $\mathcal{P}(s_t, a_t)]$ from $(s_t, a_t, s_{t+1}, a_{t+1})$
    $s_t \leftarrow s_{t+1}$
    $a_t \leftarrow a_{t+1}$
  end while
end for

2.4. Global methodology

The learning models used in this paper share three meta-parameters: the learning rate α, the exploration temperature β

and the discount factor γ. These parameters have a strong influence on the dynamics of the learning process, both at the value function level (hence the RPE signal) and at the behavior level. To fit the value of these meta-parameters to data from [34], we can use either behavioral data, dopaminergic data, or both.

If DA activity reflects a learning process that directly controls the behavior of the animal, as assumed in many model-based studies (refs), it would be desirable to fit both. Thus, we explored the parameter space of the model until we find a set of parameters that best described the learned behavior of the animal. We then extracted the trial-by-trial evolution of variables (δ, V) in this optimized model and compared the corresponding time series with DA activity to see whether they shared a common set of properties and exhibited a good fit. Fit was assessed by performing a quantitative fit based on regression from the data shown in Figure 6 in [34] and also by statistically testing whether the DA signal matches the properties outlined in [34]. These two approaches are described in Sections 2.7 and 2.8.

Of course, other processes, not under the control of the DA neurons, may also influence the animal's behavior. In this case, the behavioral dynamics may be partly disconnected from the variations of DA activity. To allow for this possibility, we also fit the parameters of each model only to neural activity, using a fixed behavioral policy extracted from the learning curves of the animals. The simulated animals just learn the critic part of their model, using this fixed policy. Section 2.6 describes the procedure employed in this case.

2.5. Fitting both DA signal and Behavior

To reproduce the behavioral results of Roesch et al. [34], 50 simulated agents learned the contingency of each block of a session during 90 trials, using the previously presented algorithms. Each odor was presented once every three trials. Thus, in one block, 30 trials of each odor were presented to the model, which is on average the number of trials per blocks that the rats experienced in the experiment. The behavior of each model described in the previous section was computed as the number of left choices during each free choice trial (odor 3) and was compared to that of real rats. More precisely, the behavior of each model was recorded during the last 15 trials of the first block and during 30 trials in the next block (the behavior is recorded after the block change between blocks 1 and 2 and between blocks 3 and 4), averaged over 50 agents experiencing a session. This simulated behavior was then comparable to the behavior reported in Roesch et al.'s experiment.

Each algorithm has three parameters α, β and γ (see Section 2.3) that influence behavior. For each algorithm, we performed a grid search testing all the combinations of the following values for these parameters:

- α: from 0.1 to 0.9 with 0.05 steps (lower values between 1E-4 and 0.1 have also been tested),
- β: from 0.1 to 1 with 0.05 steps; we also tested 1.5 and 2,
- γ: from 0.1 to 0.9 with 0.1 steps; we also tested 0.99.

The obtained results were compared, for each parameter set, to those of Roesch et al. [34] by minimizing the distance between simulated and experimental behavior. The points in the curves of Roesch et al. [34] in the *size* case and the *delay* case are the percentage of left choices during each trial of the neural recording sessions (see figure 1B-C). We searched for the set of parameters that would optimize the match between our models' and these data in both cases (size and delay).

2.6. Fitting DA signal with a fixed behavioral policy

In the second part of this work we investigated the ability of the previously described learning rule to reproduce the DA activity without requiring that the algorithms also fit the animal's behavior. The action is thus not chosen with a *softMax* based on the value learned by the algorithms, as described in Algorithm 1, but with a dedicated function *chooseAction* built to reproduce the behavior of the rats in [34]. For each block, we associated to each trial of this block the observed probability to choose the 'left' action. This probability is extracted from the behavior of rats during a reversal. Thus the probability of performing the left action in a specific simulated trial is defined by the frequency with which actual rats chose the left action in the same trial (see Figure 1B-C for the frequency of left action during the reversals). Outside the trials where the behavioral data is fully accessible (before the last 15 trials in the first block of the delay and size case), we assumed that the animal chose the most rewarding action in 70% of trials during free choice on average [34]. During forced choice trials (odor 1 and odor 2), the animals learned to perform the best action on nearly all trials, so the agent chooses the best action 99% of the time in these trials. All other fitting procedures (quantitative and qualitative) are the same as for the case with the behavioral constraints.

2.7. Quantitatively fitting DA activity

As DA activity recorded in this experiment shows a large phasic response to the expected reward despite stabilization of the learned behavior (i.e. behavioral convergence), we hypothesized that this signal could be better reproduced by a value function (i.e. the sum of the immediate reward, r_t, plus future expected reward $V(s_t)$ for Actor-Critic, $Q(s_t, a_t)$ for Sarsa and $\max_a(Q(s_t, a))$ for Q-learning), instead of a pure classisal RPE. Note that we did not modify the internal operations of the algorithms, we only searched for the combination of the internal variables of these algorithms that would best explain the neuron activity. To test this hypothesis, for every simulation (e.g. set of parameters), we tested the ability of 10 different mixtures, M_w with w $\in [0, 1]$ with 0.1 steps, of value function and RPE to reproduce previously recorded DA activity. We defined:

$$\begin{aligned} M_w(t) &= wValue(t) + (1-w)RPE(t-1) \\ &= w[r_t + \gamma V(s_t)] + (1-w)[r_t + \gamma V(s_t) - V(s_{t-1})] \\ &= r_t + \gamma V(s_t) + (1-w)V(s_{t-1}) \end{aligned}$$

This equation illustrates the mixture for the Actor-Critic. The future expected reward $V(s_t)$ is replaced by $\max_a(Q(s_t, a))$

for Q-LEARNING and by $Q(s_t, a_t)$ for SARSA. Of course the RPE signal actually incorporates the value signal since the RPE is the difference between the current value and the previous prediction of the value. Thus the mixture that we used as a regressor for DA activity can be interpreted as a distorted or optimistic RPE where the negative part – i.e. the previous prediction of value – is underweighted.

In order to compare the DA activity with a mixture computed by the different algorithms, we fitted three states of our MDP with experimental data, corresponding to the three previously presented salient events (see Figure 2B): the nosepoke, the perception of the odor and the reward delivery or omission.

As DA activity and simulated mixtures do not share a common scale nor the same baseline, we authorized a linear transformation of the simulated signal to fit DA activity recorded in vivo. We minimized the difference between both values with least squares (LS) by minimizing the error $e = \|(aM_s + b) - DA_s\|^2$ where DA_s is the experimental DA activity averaged over the trials after the performance of the rat went above 50% in state s and M_s is the average mixture computed in s during the trials of a block.

Thus we have: $M_w(s) = \frac{1}{n}\Sigma_{e=0}^{n} M_w^e(s)$, where n is the number of considered trials and $M_w^s(e)$ is the mixture computed from the e^{th} trial in s. The (a, b) pair is determined with the LS method. The error reported in this study, noted LS error, is the error obtained with a mixture averaged over all simulated agents. This LS error gives us a quantitative evaluation of the ability of our model to reproduce DA activity.

2.8. Qualitatively fitting DA activity

In the original study, the authors compared DA activity at reward delivery and omission between early and late trials of each block. This analysis was performed to show that, as in previous work [2, 13, 18, 36, 37, 40], DA activity was significantly lower during the first omission trials than during the last omission trials of a block and was significantly higher during the first delivery trials than during the last delivery trials. In addition, they performed t-tests on DA activity at the time of the odor, comparing trials where the animal chooses the best option with trials where the animal chooses the less attractive one under all conditions. The initial question was to see whether the DA signal would be influenced by the future action of the animal or not. They found that there was no statistical difference in DA activity depending on the chosen action in free choice trials. Moreover, DA activity recorded during free choice was not statistically different from DA activity recorded during forced choices that led to the best option in each block, but it was statistically different from DA activity recorded during forced choices that led to the worst option.

Hence, to further assess qualitatively the ability of the three models to reproduce DA activity, we also included a statistical analysis of the activity predicted by the models at the time of the odor in the different conditions (free choice long/short and small/big and forced choice long/short and small/big). This analysis involved 12 different statistical tests (see Table 1), in order to reproduce the previously observed pattern of activity

	free Big	free Small	forced Big	forced Small
free Big	X	=	=	>
free Small	=	X	=	>
forced Big	=	=	X	>
forced Small	<	<	<	X

	free Short	free Long	forced Short	forced Long
free Short	X	=	=	>
free Long	=	X	=	>
forced Short	=	=	X	>
forced Long	<	<	<	X

Table 1: Description and results of the different t-tests used in the original study. =: indicates that $p > 0.05$ and the data are not statistically different. > or <: indicate that $p < 0.05$ and the row data is significantly superior or inferior respectively to the column data.

at the time of the odor. Two additional tests were added to assess the evolution of the simulated activity during omission and delivery after a reversal as mentioned previously. As in the original study, we used t-tests to determine if the activity in these different cases was the same or not. If the p-value is larger than 0.05 then we did not reject the null hypothesis of identical average activity, otherwise we considered the activity of both cases to be statistically different.

To take these results into account in our study, we attributed to each model a Statistical Tests (ST) score, which was defined as the number of statistical tests that a given model satisfied out of the tests performed.

Using this method, we could analyzed both the precise pattern of activity predicted by our models at the time of the choice and the evolution of this activity during the omission and delivery trials.

3. Results

3.1. Fitting the rats behavior

In Roesch and colleagues' work, rats learned to choose more often the well associated with the best available option (big reward in the *size* case and short reward in the *delay* case). In free choice trials, after the contingency of a block was learned, rats chose the best option 75-80% of the time. Fifteen to twenty trials were necessary for animals to adapt to the new contingency after a reversal. In forced choice trials, however, rats quickly learned to choose the well leading to the better reward [34].

We looked at the ability of previously presented TD learning algorithms to reproduce the behavioral adaptation of rats during free choice, which required from them to change their behavior after each reversal to maximize their reward. We tested different combinations of learning rate α, temperature β for a correct exploration/exploitation trade-off, and the discount factor γ required to generate with different algorithms (SARSA, Q-LEARNING and ACTOR-CRITIC) a behavior close to the one produced by the rats in the Roesch et al. experiment.

Figures 3 A, D and G report the LS error between experimental and simulated behavior for Q-LEARNING, SARSA and ACTOR-CRITIC as a function of the different parameters. These results show that Q-LEARNING and SARSA minimize the error in a specific

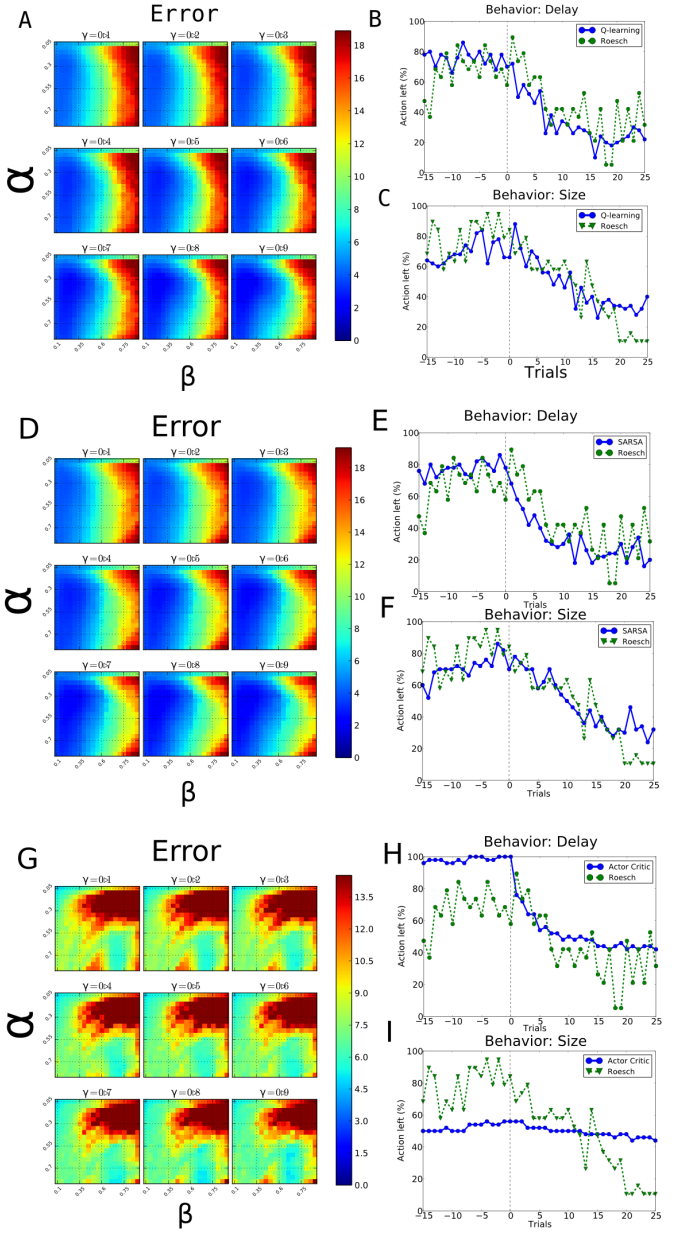


Figure 3: Reproduction of experimental behavior with Q-LEARNING, SARSA and ACTOR-CRITIC. A, D and G. LS error in function of the parameters: the learning rate α; the temperature β and the discount factor γ for respectively Q-LEARNING, SARSA and ACTOR-CRITIC. B, E and H. Best behavioral fit for the *delay* for respectively Q-LEARNING, SARSA and ACTOR-CRITIC. C, F and I. Best behavioral fit for the *size* case for respectively Q-LEARNING, SARSA and ACTOR-CRITIC.

region of the parameters space. Indeed, with a learning rate α and β around 0.3, the behavior seems to be reproduced with a very low error (see Figures 3 A and D). A large γ also appears to helps the algorithms to better fit the behavior. Q-LEARNING and SARSA minimize the distance between their behavior and rats' behavior with the same parameters set: $\alpha = 0.35$, $\beta = 0.25$ and $\gamma = 0.9$. With this parameter set, both algorithms reproduced the behavioral adaptation well, choosing the 'left' action, before block change 70-80% of the time and at the end of the post reversal block, around 20-30%; which match rats' behavior in this task (see Figures 3 B-C and E-F for the best fit of respectively Q-LEARNING and SARSA). These results highlight the strong similarity between these two algorithms. Even if the calculation of the RPE is slightly different, both algorithms build their policy and calculate the RPE signal based on Q-value (see Methods for more details), and show no significant behavioral differences. However, the ACTOR-CRITIC model shows a much different sensitivity to the parameters compared to Q-LEARNING and SARSA. Indeed for this model, the error is only minimized with a much larger learning rate, α (around 0.8) and a much larger β parameter (see Figures 3 G) compared to what is required to reproduce rats behavior for Q-LEARNING and SARSA. But even when considering the parameters that produce the best fit ($\alpha = 0.85$, $\beta = 0.7$ and $\gamma = 0.9$), the generated behavior still does not produce a satisfying fit. Indeed, while in the first delay block, the behavior converged to 100% of 'left' actions, in the following blocks the averaged behavior seems to be stuck at 50% of 'left' actions, which does not approach the behavioral adaptation of rats.

Other studies [3, 28] show that ACTOR-CRITIC has a limited ability to reproduce animal's behavior during reversals such has the one described in [34]. Unlike Q-LEARNING or SARSA, ACTOR-CRITIC needs to learn both a value function that encodes the future expected reward knowing the current state, $V(s_t)$ and a value $P(s_t, a_t)$ from which the policy is inferred. It seems that this architecture needs more time to adapt to a block change and is less suited to perform multiple reversals. This architecture tends to create optimal behavior (see Figure 3 H) before any reversal by creating a larger difference in the P-value than in the Q-value calculated by SARSA and Q-LEARNING. As the policy is inferred from P-values, if there is a large difference between them, the probability to choose the best action is increased.

Given a policy π, we have:

$$V(o3) = \pi(o3,'left')[r_{left}+V(r_{left})]+\pi(o3,'right')[r_{right}+V(r_{right})]$$

Here, o3 stands for odor 3 and indicates the state where models receive the free choice cue. V(o3) is thus inferior to the value of the best option and superior to the value of the worst option if the policy π is stochastic (if the policy always chooses one or the other action then V will converge to the value of this action, a, which is $Q(s,a)$). Hence, when the worst action is chosen, the RPE is negative: $\delta_{worst} = r_{worst}+V(r_{worst})-V(s) < 0$ thus $P(s, worst)$ is decreased according to the updates rule of ACTOR-CRITIC (see method for more details). We have the opposite effect for the best action. Thus the P-value for the best and worst option does not converge to the actual future expected reward, but while the policy is stochastic these values diverge forming an important gap between them. This important difference makes a reversal more difficult to learn because the models used here need to unlearn previous values before being able to learn the new ones.

However, during these simulations we forced the learning rate of the critic to be the same as the learning rate of the actor, which is not consistent with theoretical studies that suggest that the learning rate of the critic should be lower than the learning rate of the actor [4, 23]. To test further the ability of ACTOR-CRITIC to reproduce rats' behavior in this task, we conducted

additional simulations to test whether a different tuning of the parameters, and especially if two different learning rates for the actor and the critic, could generate a better fit. However, the best fit happened with a low learning rate (for the actor and/or the critic) and/or a low temperature which all result in a very exploratory policy inconsistent with rats' behavior.

In summary, our results show that Q-LEARNING and SARSA accurately reproduce rats' behavior during the *delay* and *size* reversal with the same meta-parameter set, whereas the ACTOR-CRITIC is unable to reproduce this behavior due to its different architecture.

3.2. Fitting DA activity under behavioral constraint

Based on the parameters obtained from the behavioral fit, we investigated whether simulations using the RPE or a mixture of value and RPE could match the DA activity observed in rats. If DA activity reflects the RPE signal of the algorithm by which rats learn the task, algorithms tuned to fit the behavior should display the same pattern of activity as observed in the responses of the DA neurons. To evaluate the ability of a signal to reproduce DA activity, we had two criteria: 1) the ST score based on the ability of the signal to reproduce the pattern of DA activity recorded at the time of the odor and during reward omission and delivery through learning (see Methods for more details) and 2) the fitting LS error obtained by minimizing the distance between the DA signal and a linear transformation of the simulated signal.

Our results show that the RPE signal simulated with either Q-LEARNING or SARSA converged too much and could not explain the high DA response to the reward in the experimental data (see Figures 4 A and C). The best fit with a pure RPE calculated by SARSA or Q-LEARNING consisted of an almost flat signal indicating that the algorithms learned to predict the outcome fully and thus do not make any prediction error at reward. This very low response at the time of the reward is not consistent with recorded DA activity under the same conditions. However for both Q-LEARNING and SARSA, a pure RPE signal obtains a better ST score (e.g. the number of statistical tests the model could satisfy) than any other mixture (see Figure 5 A and B). Q-LEARNING gets a ST score of 12 for a pure RPE signal and SARSA gets a ST score of 10. This indicates that, consistent with the interpretation of [34], for these algorithms and in the case of parameters constrained by the behavior, a pure RPE signal can better reproduce the pattern of activity at the time of choice and that both RPE signals could reproduce the evolution of DA activity during reward omission and delivery. A mixture with too much weight on the value function could not reproduce such an evolution (see Figure 5).

The simulated value on the other hand can better reproduce the global pattern of DA activity (see Figure 4 B and D) for SARSA and Q-LEARNING by reproducing the growing activity as the simulated agent gets closer to the reward. The LS error is thus smaller for the value function than for an RPE signal or any other mixture. More globally, the ST score of the value function is lower than the one of the RPE signal. However the value function cannot reproduce the evolution of DA activity observed experimentally during reward omission and delivery.

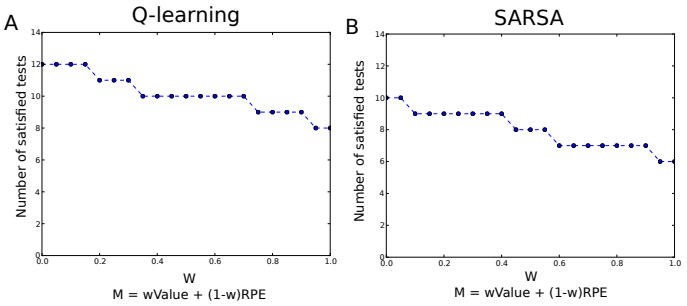


Figure 5: Number of valid tests (ST score) in function of the w weight of the mixture for Q-LEARNING (A) and SARSA (B) when fitting under behavioral constraint. If w=1 then the mixture represents the sum of the immediate plus future expected reward (i.e. a pure value function) and if w=0 then the mixture represents a pure RPE signal.

More generally, the smallest LS error was observed when considering a pure value signal while the highest ST score was observed with a pure RPE signal (see Figure 5). Moreover, Q-LEARNING seems to be better suited to reproduce the observed data, since it predicts activity that is not action dependent on free choice trials (see Figure 4 A and B). By contrast, SARSA predicts divergent signals on free (and forced) choice trials, depending on the chosen action (see Figure 4 C and D).

Our results show that the DA signal is neither a pure RPE signal nor a pure value signal. Although Q-LEARNING obtains better results than SARSA, consistent with the interpretation of Roesch et al. 2007 [34], when fitted on the rat's behavior, the RPE signal of both algorithms converge too much to reproduce the observed pattern of DA activity at the time of reward. Thus no mixture fully reproduces the pattern of activity at the time of the odor,the high response to the reward, andthe evolution of the signal during the omission and delivery trials.

One possible explanation is that the DA signal recorded here only partly reflects the learning process that underlies the observed behavior. This hypothesis is based on numerous studies that suggest multiple parallel learning systems involving different parts of the basal ganglia [7, 19, 35, 43, 44]. If some learning were not under the control of the DA signal – at least that from VTA – then this would allow the signal to diverge from the prediction of models constrained by behavioral changes. Therefore, in the next part of our work we released the behavioral constraint to explore the ability of the the models to only fit DA activity.

3.3. Fitting DA activity using a fixed policy

To further investigate the nature of the information encoded by DA neurons, we released the behavioral constraint to focus on the match between the DA signal recorded in the experiment by Roesch et. al. and that generated by different simulated mixtures of value and RPE of different models.

We simulated DA activity using different models defined by: (1) the algorithms used: Q-LEARNING, SARSA or ACTOR-CRITIC; (2) the parameters of the chosen algorithm: learning rate α and

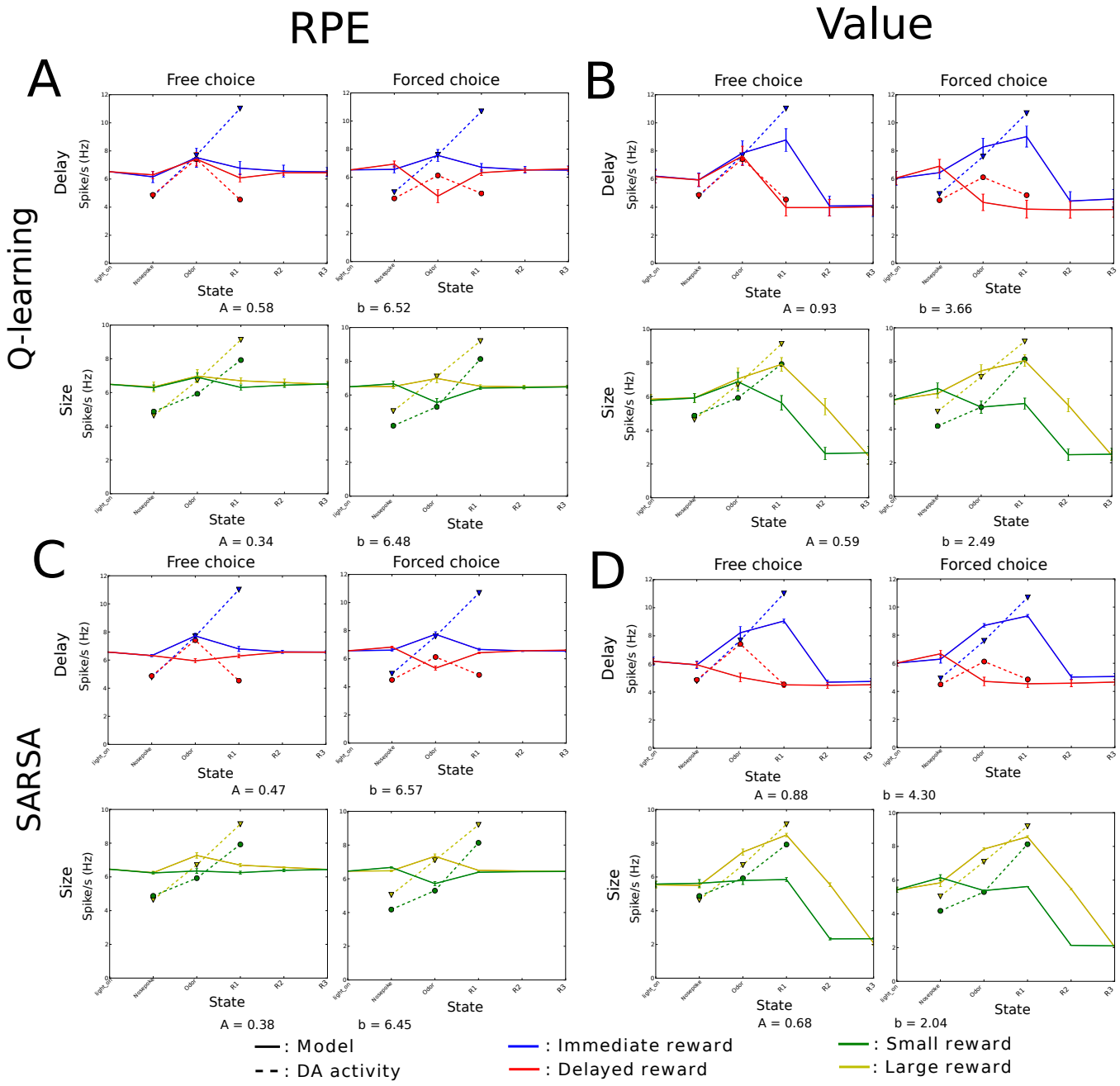


Figure 4: Reproduction of DA activity with parameters that best reproduce experimental behavior. Each subfigure illustrates the best fit of DA activity during the *delay* and *size* case for the free and forced choice. Two optimisations were performed: one for the *size* case and an other one for the *delay* case resulting in two (A,b) pairs that minimize the distance between DA signal and *A*(simulated signal) + b*. This linear transformation is necessary to compare DA signal with simulated values, which do not share the same scale nor the same baseline. The simulated signal was averaged over 50 sessions. A-B. Fit of DA activity with respectively the RPE and the value calculated by Q-LEARNING. C-D. Fit of DA activity with respectively the RPE and the value calculated by SARSA.

discount factor γ (β being no longer needed since the policy is fixed); (3) the mixture parameter, w, that defines the mixture (if w = 1 the signal is a pure value signal; if w = 0 it is a pure RPE signal). As in the previous section, we attributed a score to each model according to the number of statistical tests it could satisfy (ST score), and an error based on the fitting error it makes with observed DA (LS error). Figure 6 A, B and C report the results from models that were able to reproduce the evolution of the activity during omission and delivery. Unexpectedly, once the behavioral constraint was relaxed, only the ACTOR-CRITIC model was able to reproduce all 14 statistical tests or achieve 13 correct tests out of 14 with a lower LS error. Q-LEARNING reproduced 12 tests and SARSA only 10. The fact that SARSA was less suited to reproduce DA activity at the time of the choice, confirms that at this time in this particular behavioral setting, VTA DA did not encode any information based on the chosen

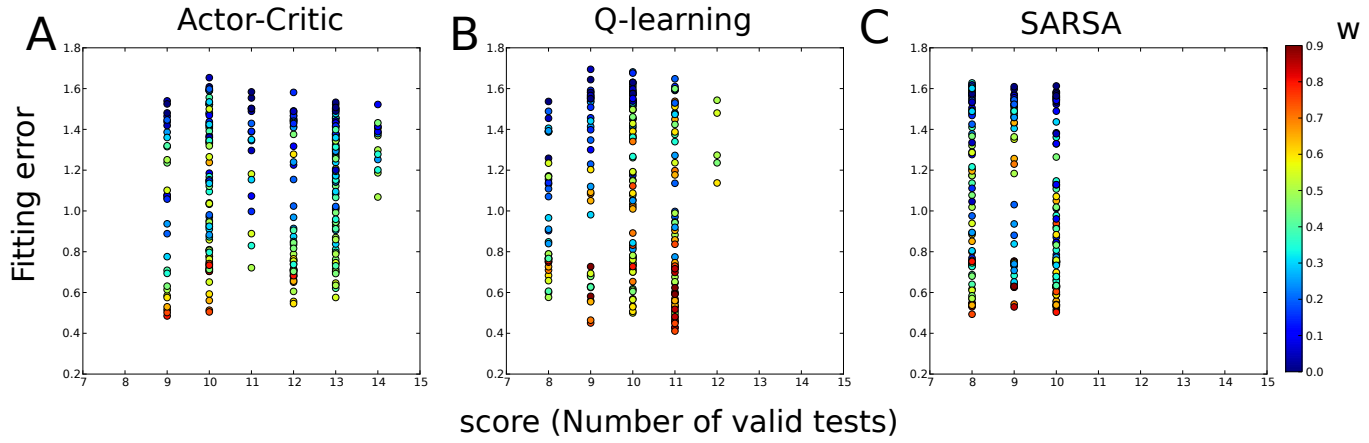


Figure 6: Performance of the different models to reproduce statistical tests and fitting error. Each figure displays the ST score and LS error of each model generated by one of the three studied algorithms: A. Actor-Critic, B. Q-learning and C. Sarsa. Each dot in the figures represents one model. The color of the dot represents the w weight of the mixture used to calculate its ST score and LS error. Blue dots represent models with a pure RPE signal (w = 0) and red dots represent models with 90% value (w = 0.9). Models that could not reproduce the evolution of DA activity during reward and omission have been excluded. Hence no model with pure Value are plotted in this figure.

action. Unexpectedly though, the Actor-Critic received a better overall score than Q-learning. Hence, based on the analysis of the pattern of DA activity at the time of choice, the Actor-Critic was the best candidate to explain DA activity. This suggests that the signal encoded by putative DA neurons represents information based on the value of the current state rather than on the action value, as would be predicted by both Q-learning and Sarsa.

However Actor-Critic models that reproduced every test show a large LS error when fitting DA activity. A good compromise was the Actor-Critic model that best fit DA activity while reproducing 13 tests out of 14 (see Figure 7). The parameters used by this model are $\alpha = 1$, $\gamma = 0.3$ and w = 0.55. Thus the best model of DA activity resulted from a balanced mixture of both current value and RPE. Consistently with previous results, we can see that models with a high w, indicating an almost pure value, seems to get a low LS error but a lower ST score as well (red dots in Figures 6 A-C). However models with a low or null w (pure RPE signal), have a tendency to get a large LS error but can get a good ST score (blue dots in Figures 6 A-C).

A feature of the Actor-Critic model that could help better reproduce the signal is that it calculates the RPE based on state values rather than on action values. As a result, even though the learning rate is large, there is still a remaining RPE signal at the time of the reward. Indeed, as the policy is based on the behavior of the animal, in free choice trials, the less attractive reward is chosen about 30% of the time. As V represents the average future expected reward when the worst (respectively best) option is chosen, there is a negative (respectively positive) RPE. As Q-learning and Sarsa calculate the RPE based on action values, there is no remaining RPE signal after convergence of the Q-value.

To summarize this study, in the case with no behavioral constraint, the best model we can propose to reproduce DA activity in this task is a mixture of RPE and value (w = 0.55), calculated by Actor-Critic.

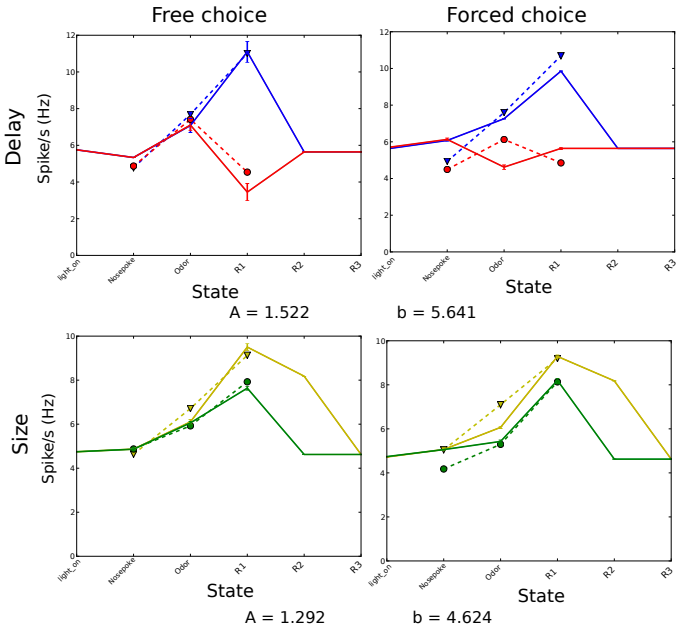


Figure 7: Reproduction of DA activity with the best simulated model obtained with Actor-Critic. The ST score of this model is 13 and gets the lower LS error for a model with that ST score. The parameters of the model are: $\alpha = 1$, $\gamma = 0.3$ and w = 0.55.

4. Discussion

In this study, we quantitatively compared the ability of different TD learning algorithms to reproduce DA activity recorded in the multi-choice task in Roesch et al. 2007 [34].

Our starting hypothesis, based on DA recording in passive monkeys [37], was that DA neurons activity would reflect an RPE pattern [2, 13, 18, 40], compatible with the reinforcement signal used in TD learning algorithms [10, 37, 39]. We further assumed that learning would be dependent upon or constrained

by this dopaminergic teaching signal. We also specifically assessed the role of future action in determining the signal.

In most previous studies, DA activity has been recorded during Pavlovian conditioning before and after extended training [2, 13, 29, 37]. In such settings, it is difficult to compare the convergence of the animals' behavior and DA signaling. Moreover, the experimental context was easily predicted in these studies by contrast with the experimental set-up used which included frequent reversals, which forced rats to constantly monitor and adapt their behavior to new contingencies. In standard reinforcement learning theory, learning happens exclusively from the RPE signal, suggesting a conjoint evolution between behavior and RPE. Hence, in the first part of our work, we tuned the studied algorithms with parameters that reproduce the observed behavior. If DA activity does reflect an RPE and if that RPE is the sole arbiter of learning, then the RPE calculated in the simulation constrained by the animals' behavior should be best able to reproduce the observed DA activity. Interestingly the Actor-Critic was unable to reproduce the change in behavior (see Figure 3 G,H and I). This inability seems to be due to the multiple reversals used in this task and is consistent with a study that found that Actor-Critic is less suited to reproduce animals' behavior after a reversal [28].

Further, while Sarsa nor Q-learning were able to reproduce the animals' behavior, neither produced an RPE signal consistent with recorded DA activity. Instead both exhibited too much convergence in contrast with the large phasic DA response to the reward observed experimentally (see Figure 4 A and C). The value function – defined as the sum of immediate plus future expected reward – generated a signal that was quantitatively better suited to fit this activity than a pure RPE signal (see Figure 4 B and D). However this function was unable to reproduce the pattern of activity observed at the time of the choice of the animal (a value function cannot explain the evolution of DA activity during omission and delivery).

These results underline some strong limitations of the standard *TD* learning algorithms to reproduce both DA activity, interpreted as a mixture of value and RPE, and rats' behavioral adaptation. They suggest that the information carried out by DA might not be fully related to behavioral changes as assumed by reinforcement learning algorithms. This raises a question about the link between both behavioral convergence and the convergence of the information encoded by DA neurons. It would be interesting to look at the conjoint evolution of both from early trials to late in training in a stable environment to see whether the discrepancy observed here is due to the instability introduced by frequent reversals. Our results would predict different DA activity after full convergence depending on how we interpret the mixture function used here. Indeed, we can interpret our mixture in two different ways. We can consider that we strengthened the value part of the RPE signal or we weakened the prediction signal of the RPE signal. If the value signal is actually over represented in DA neurons it would imply that even after extensive training, these neurons should still, to some extent, respond to expected reward. On the contrary, if the prediction signal is weaker in early trials or in an uncertain environment, we can imagine that after extensive training the prediction signal would increase until the global signal sent by DA neurons would converge to a pure RPE signal, and these neurons would then stop responding to the reward. This would suggest that the value and prediction part of the RPE are not learned at the same speed. These two interpretations would predict different activity after extensive training but both of them imply that, at least during early trials, the signal encoded by VTA DA neurons does not encode an RPE or mixture consistent with behavioral adaptation.

The apparent dissociation between DA activity and behavioral adaptation that we found is also consistent with the idea that behavior depends on multiple parallel learning systems, some operating independent of DA error signaling. This is perhaps not a surprising notion. While reinforcement learning models assume that behavior is only driven by one process guided by RPE, many neurobiological studies have suggested that behavior is the results of complex interaction between parallel neural systems One popular model has proposed that different cortical and subcortical circuits control habitual versus goal directed behavior [1, 42]. Such a dissociation has been modeled by multiple studies [7, 9, 22]. Our results suggest that VTA DA signal might only be a part of one learning system, with the behavior resulting from an interaction between this system and other parallel systems. These systems might be dependent on DA signals not assessed here, perhaps from SNc for example, or they might also be DA independent, as suggested by the work of Flagel et al. [14, 26] showing that some rats are more or less dependent from DA to learn a task.

Thus, in the second part of this work, we assumed that DA activity might be not fully correlated with behavioral adaptation and we released the behavioral constraint over the parameters to focus on the decoding of the information encoded by DA neurons. The question we wanted to answer was whether DA activity could reflects an RPE with a slow convergence, which would lead to a maintained RPE at the time of the reward and would explain the high DA phasic response; or if only a value function or a mixture of both informations would better fit this activity. Our results suggest that a mixture of value and RPE calculated by Actor-Critic is actually the best model to reproduce the DA activity observed in this task (see Figure 6 and 7) and that no pure RPE signal could reproduce such activity. Even though the task implies lots of reversals, which could to some extent explain why DA signal could still be strong at the time of the reward, some rewards are given in many concurrent blocks. Indeed, when looking at the progress over one session (see Figure 1 A), in the right well from the block 2 to 4 at least one reward is given (for the short, small and big rewards). In that condition, *TD* learning algorithms have plenty of time to learn to predict this first reward and it was thus not possible to reproduce the high DA response to this reward with an RPE signal even with a low learning rate. The fact that no pure RPE signal could reproduce DA activity, even without any behavioral constraint, is unexpected because it goes against the widely accepted hypothesis that DA neurons activity reflects an RPE signal [2, 13, 18, 33, 37, 40]. However a previous study [11], also challenged this common hypothesis by comparing DA activity with the long term value of multiple future rewards. Moreover

in absence of previous information RPE and value are similar making the difference between both information very small in some contexts. As a conclusion, the exact nature of the information encoded by DA neurons is still unclear.

We can partly answer our initial question which is: which *TD* learning algorithms can better reproduce DA activity in this task? Our results here clearly indicates that only a non action dependent signal can reproduce this activity as claimed in the original study. We also found that Actor-Critic, when looking only at the convergence of the mixture signal, slightly better reproduce DA activity than Q-learning. We suppose that this advantage is due to the fact that the calculation of the RPE in Actor-Critic is only based on the current state without any action value. However one can note that during the learning process $V(s)$ often converges to the value of the best option, making the difference between the two algorithms very small after convergence.

But if our results validate that the information encoded by DA neurons is not action dependent, then how can we explain that Morris et al. [31], found an action dependent activity ? This question might be answered by the differences in the setup of the tasks, or by the fact that Morris et al. [31] used monkeys while rats were used in Roesch et al. [34]. But we can also hypothesized that the apparent contradiction between the results of these studies lies in the type of DA neurons recorded in both task. In [34], DA neurons were mostly recorded in VTA whereas in [31], DA neurons were recorded in *substantia nigra par compacta* (SNc). VTA DA neurons project preferentially to ventral part of the striatum [15–17, 21] which is considered to be the critic part of the Actor-Critic architecture of the basal ganglia [20] and our results show that the information encoded by VTA DA neurons are consistent with a signal built to update state based value ($V(s)$ in Actor-Critic) which is the critic part of the Actor-Critic architecture. The policy, on the other hand, is believed to be computed in more dorsal part of the striatum [20, 44] and receives mostly input from SNc DA neurons. Thus SNc DA neurons might encode an action dependent RPE to update the actor part of the basal ganlia (which is the dorsal part of the striatum).

Our results show that an Actor-Critic architecture was unable to adapt to many reversal, which seems in contradiction with the Actor-Critic view of the basal ganglia. But, in the standard Actor-Critic algorithm, there is only one RPE to update both the actor and the critic. The difference in DA activity between VTA and SNc DA neurons suggest that different information update the actor and critic part of the basal ganglia which would explain why rats can adapt to multiple switch when an Actor-Critic algorithm cannot. It would thus be interesting to look at the difference between VTA and SNc DA neurons activity to see if they encode different informations and possibly different types of RPE, depending on their anatomical location.

Some recent studies also show the relevance of carefully determining the projection field of the DA neurons in order to be able to interpret the information carried out by these neurons. Indeed DA projects to a broad range of areas in the brain from cortical to subcortical areas such as the amygdala and the medial prefrontal cortex. It is very likely that depending on the targeted area, DA neurons would have different role as suggested in Lammel et al. [25].

The fact that this activity can only be fitted by a mixture of two informations could imply the presence of two distinct DA neurons populations. The 17 VTA DA neurons used here were selected for their RPE coding properties: (1) they are sensitive to reward and cues that predict reward; (2) their activity tends to diminish as the reward becomes predicted; (3) their response is first inhibited by omission and tends to go back to baseline when the omission becomes predictable. Actually, these neurons can be separated in two different groups of neurons, some are more sensitive to cues predicting the reward and some to the reward itself. But both categories show on average a similar activity and are both better fitted by a mixture of RPE and value. Thus the presence of two different informations is not due to two different populations of neurons but are embodied in the same dopaminergic signal.

While questioning the validity of the RPE hypothesis, our study does not strongly contradict it. But, we can see in the literature more and more studies that question the classical RPE interpretation of the DA signal. There is growing evidence that DA neurons activity does not encode an unique signal, but seems to show an heterogeneity of response to aversive stimuli [5, 29, 38, 41], even if there is still no consensus on the heterogeneity of the response of DA neurons to aversive stimuli [12]. These different signals may be part of different circuits depending on their anatomical localisation as suggested by the work of Lammel et al. [24, 25] showing the presence of different pathways linked to reward or punishment, involving different subpopulations of DA neurons.

In summary, our work shows the limitation of the standard *TD* learning algorithms to reproduce both DA activity and behavioral adaptation and question the common RPE hypothesis by showing that the information encoded by DA neurons in a multi-choice task could not be reproduced by a pure RPE signal even without any behavioral constraint, it also shows that the Actor-Critic algorithm may have some explanatory power, at least for the ventral circuit. This shows the need to carefully investigate the link between behavioral adaptation and information encoded by DA neurons to better understand its role in learning.

References

[1] B. W Balleine and J. P O'Doherty. Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*, 35(1):48–69, January 2010.

[2] H. M Bayer and P. W Glimcher. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, 47(1):129–141, 2005.

[3] D. P Bertsekas and J. N Tsitsiklis. Neuro-dynamic programming: An overview. In *Decision and Control, 1995., Proceedings of the 34th IEEE Conference on*, volume 1, pages 560–564. IEEE, 1995.

[4] S. Bhatnagar, M. Ghavamzadeh, M. Lee, and R. S Sutton. Incremental natural actor-critic algorithms. In *Advances in neural information processing systems*, pages 105–112, 2007.

[5] F. Brischoux, S. Chakraborty, D. I Brierley, and M. a Ungless. Phasic excitation of dopamine neurons in ventral VTA by noxious stimuli. *Proceedings of the National Academy of Sciences of the United States of America*, 106(12):4894–9, March 2009.

[6] N. D Daw. Dopamine: at the intersection of reward and action. *Nat Neurosci*, 10(12):1505–1507, December 2007.

[7] N. D Daw, Y. Niv, and P. Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci*, 8(12):1704–1711, December 2005.

[8] J. J Day, J. L Jones, R M. Wightman, and R. M Carelli. Phasic nucleus accumbens dopamine release encodes effort- and delay-related costs. *Biological psychiatry*, 68(3):306–9, August 2010.

[9] L. Dollé, D. Sheynikhovich, B. Girard, R. Chavarriaga, and A. Guillot. Path planning versus cue responding: a bio-inspired model of switching between navigation strategies. *Biological cybernetics*, 103(4):299–317, October 2010.

[10] K. Doya. Reinforcement learning: Computational theory and biological mechanisms. *HFSP Journal*, 1(1):30, 2007.

[11] K. Enomoto, N. Matsumoto, S. Nakai, T. Satoh, T. K Sato, Y. Ueda, H. Inokawa, M. Haruno, and M. Kimura. Dopamine neurons learn to encode the long-term value of multiple future rewards. *PNAS*, 2011.

[12] C. D Fiorillo. Two dimensions of value: dopamine neurons represent reward but not aversiveness. *Science (New York, N.Y.)*, 341(6145):546–9, August 2013.

[13] C. D Fiorillo, P. N Tobler, and W. Schultz. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, 299(5614):1898, 2003.

[14] S. B Flagel, J. J Clark, T. E Robinson, L. Mayo, A. Czuj, I. Willuhn, C. A Akers, S. M Clinton, P. EM Phillips, and H. Akil. A selective role for dopamine in stimulus-reward learning. *Nature*, 469(7328):53–57, 2010.

[15] S. N. Haber. The primate basal ganglia: parallel and integrative networks. *Journal of Chemical Neuroanatomy*, 26(4):317–330, December 2003.

[16] S. N Haber and R. Calzavara. The cortico-basal ganglia integrative network: the role of the thalamus. *Brain research bulletin*, 78(2-3):69–74, February 2009.

[17] S. N Haber, J. L Fudge, and N. R McFarland. Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 20(6):2369–82, March 2000.

[18] J. R Hollerman and W. Schultz. Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat Neurosci*, 1(4):304–309, 1998.

[19] M. Ito and K. Doya. Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit. *Current opinion in neurobiology*, 21(3):368–73, June 2011.

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

[21] D. Joel and I. Weiner. Commentary the connections of the dopaminergic system with the striatum in rats and primates: an analysis with respect to the functional and compartmental organization of the striatum. *Neuroscience*, 96(3):451–474, 2000.

[22] M. Keramati, A. Dezfouli, and P. Piray. Speed/Accuracy Trade-Off between the habitual and the Goal-Directed processes. *PLoS Comput Biol*, 7(5):e1002055, May 2011.

[23] V. R Konda and J. N Tsitsiklis. Actor-critic algorithms. In *NIPS*, pages 1008–1014. Citeseer, 1999.

[24] S. Lammel, D. I Ion, J. Roeper, and R. C Malenka. Projection-specific modulation of dopamine neuron synapses by aversive and rewarding stimuli. *Neuron*, 70(5):855–62, June 2011.

[25] S. Lammel, B. K. Lim, C. Ran, K. W. Huang, M. J Betley, K. M Tye, K. Deisseroth, and R. C Malenka. Input-specific control of reward and aversion in the ventral tegmental area. *Nature*, October 2012.

[26] F. Lesaint, O. Sigaud, S. B. Flagel, T. E. Robinson, and M. Khamassi. Modelling Individual Differences in the Form of Pavlovian Conditioned Approach Responses: A Dual Learning Systems Approach with Factored Representations. *PLoS Computational Biology*, 2014.

[27] T. Ljungberg, P. Apicella, and W. Schultz. Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of Neurophysiology*, 67(1):145 –163, January 1992.

[28] K. Lloyd, N. Becker, M. W Jones, and R. Bogacz. Learning to use working memory: a reinforcement learning gating model of rule acquisition in rats. *Frontiers in computational neuroscience*, 6(October):87, January 2012.

[29] M. Matsumoto and O. Hikosaka. Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature*, 459(7248):837–841, 2009.

[30] J. Mirenowicz and W. Schultz. Importance of unpredictability for reward responses in primate dopamine neurons. *Journal of Neurophysiology*, 72(2):1024 –1027, 1994.

[31] G. Morris, A. Nevet, D. Arkadir, E. Vaadia, and H. Bergman. Midbrain dopamine neurons encode decisions for future action. *Nat Neurosci*, 9(8):1057–1063, 2006.

[32] Y. Niv, N. D Daw, and P. Dayan. Choice values. *Nature neuroscience*, 9(8):987–988, 2006.

[33] Y. Niv, M.O. Duff, and P. Dayan. Dopamine, uncertainty and td learning. *Behavioral and Brain Functions*, 1:6:1–9, 2005.

[34] M. R Roesch, D. J Calu, and G. Schoenbaum. Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nat Neurosci*, 10(12):1615–1624, December 2007.

[35] K. Samejima and K. Doya. Multiple representations of belief states and action values in corticobasal ganglia loops. *Annals of the New York Academy of Sciences*, 1104:213–28, May 2007.

[36] W. Schultz. Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, 80(1):1 –27, July 1998.

[37] W. Schultz, P. Dayan, and P. R. Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593 –1599, March 1997.

[38] W. Schultz and R. Romo. Responses of nigrostriatal dopamine neurons to high-intensity somatosensory stimulation in the anesthetized monkey. *Journal of Neurophysiology*, 57(1):201–217, 1987.

[39] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, March 1998.

[40] S. C Tanaka, K. Doya, G. Okada, K. Ueda, Y. Okamoto, and S. Yamawaki. Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature Neuroscience*, 7(8):887–893, 2004.

[41] D. V Wang and J. Z Tsien. Convergent processing of both positive and negative motivational signals by the VTA dopamine neuronal populations. *PloS one*, 6(2):e17047, January 2011.

[42] H. H Yin and B. J Knowlton. The role of the basal ganglia in habit formation. *Nature reviews. Neuroscience*, 7(6):464–76, June 2006.

[43] H. H Yin, B. J Knowlton, and B. W Balleine. Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience*, 19(August 2003):181–189, 2004.

[44] H. H Yin, S. B Ostlund, B. J Knowlton, and B. W Balleine. The role of the dorsomedial striatum in instrumental conditioning. *The European journal of neuroscience*, 22(2):513–23, July 2005.

Robotic implementations of learning models

4

Contents

This chapter presents robotic implementations of neuro-inspired models of the coordination of MB and MF RL. The work is presented under the form of two papers, one published in a journal (Caluwaerts et al. 2012b), the other in the proceedings of an international conference (Renaudo et al. 2014), aiming at testing the ability of such neurocomputational models to improve robots' flexibility and adaptivity in real-world application, and in return getting new insights into the properties of these computational models when tested in these more realistic conditions.

The first one has been mainly performed by a previously supervised Master student, Ken Caluwaerts, and shows that the coordination of MB and MF learning systems for multiple-strategy-based navigation enables the robot to autonomously learn to exploit the advantages of each strategy in each subpart of the environment. The model autonomously learns that the MB strategy is more efficient to plan movements towards the goal when the robot is located far from it, but it is inefficient close to the goal because it allows only coarse movements. In contrast, the MF system was found to be more efficient to control the robot for fine-grained movements close to the goal. In locations where neither systems were efficients, the model autonomously learned to prefer a random exploration strategy. Finally, the model could learn contexts in which different location-strategies associations (*i.e.* tasksets) are learned, and could quickly restore the corresponding associative memory as soon as a previously experienced context was recognized, a process which is closer to the hypothesized role of the Hippocampus-Prefrontal Cortex network in Cognitive Control during this type of set-shifting tasks (Peyrache et al. 2009, Benchenane et al. 2010).

The second one presents the work of PhD student Erwan Renaudo and shows that the coordination of MB and MF RL also enables to exploit the advantages of each system during a habit learning task in a humanoid robot. The robot has to learn to sequentially push cubes moving on a treadmill, while minimizing computation cost : each access to the robot's camera has a cost ; each arm movement has a cost (so that the robot do not trivially move its arm all the time until touching a cube by chance) ; letting a cube fall has also a cost. When the task conditions are stable, the MB system finds an optimal behavioral strategy quicker than the MF system. It is thus initially preferred by the system coordination module. But since the MB system has a higher computational cost than the MF system, the latter later takes over the behavior once it is able to solve the task through the execution of learned habits. When a task changed is imposed, the model autonomously learns to favor again the MB system which adapts more rapidly and later let new habits be acquired.

Both robotic studies shows that MB and MF systems do not behave exactly as expected by previous computational model simulations when they are interacting during embodied real-world applications.

4.1 Parallel navigation strategies in a rat robot

4.1.1 Caluwaerts, Staffa, N'Guyen, Grand, Dollé, Favre-Félix, Girard, Khamassi (2012) Bioinspiration & Biomimetics

Bioinspir. Biomim. **7** (2012) 025009 (29pp) doi:10.1088/1748-3182/7/2/025009

A biologically inspired meta-control navigation system for the Psikharpax rat robot

K Caluwaerts[1,2,3], M Staffa[1,2,4], S N'Guyen[1,2,5], C Grand[1,2], L Dollé[1,2], A Favre-Félix[1,2], B Girard[1,2] and M Khamassi[1,2]

[1] Institut des Systèmes Intelligents et de Robotique (ISIR), Université Pierre et Marie Curie, 4 place Jussieu, 75005 Paris, France
[2] UMR7222, Centre National de la Recherche Scientifique, 75005 Paris, France
[3] Reservoir Lab, Electronics and Information Systems (ELIS) Department, Ghent University, Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium
[4] Dipartimento di Informatica e Sistemistica, Università degli Studi di Napoli Federico II, Via Claudio 21, 80125 Naples, Italy
[5] Brain Vision Systems, 75013 Paris, France

E-mail: ken.caluwaerts@ugent.be and mehdi.khamassi@isir.upmc.fr

Received 28 July 2011
Accepted for publication 9 December 2011
Published 22 May 2012
Online at stacks.iop.org/BB/7/025009

Abstract
A biologically inspired navigation system for the mobile rat-like robot named Psikharpax is presented, allowing for self-localization and autonomous navigation in an initially unknown environment. The ability of parts of the model (e.g. the strategy selection mechanism) to reproduce rat behavioral data in various maze tasks has been validated before in simulations. But the capacity of the model to work on a real robot platform had not been tested. This paper presents our work on the implementation on the Psikharpax robot of two independent navigation strategies (a place-based *planning* strategy and a cue-guided *taxon* strategy) and a strategy selection meta-controller. We show how our robot can memorize which was the optimal strategy in each situation, by means of a reinforcement learning algorithm. Moreover, a context detector enables the controller to quickly adapt to changes in the environment—recognized as new contexts—and to restore previously acquired strategy preferences when a previously experienced context is recognized. This produces adaptivity closer to rat behavioral performance and constitutes a computational proposition of the role of the rat prefrontal cortex in strategy shifting. Moreover, such a brain-inspired meta-controller may provide an advancement for learning architectures in robotics.

(Some figures may appear in colour only in the online journal)

1. Introduction

1.1. The Psikharpax robot

The Psikharpax robot [1] is designed as an *artificial rat*, a robotic platform built to integrate computational models of the rat's decision, learning, motivational and navigation circuits. It is used for two purposes: as a tool to contribute to neuroscience by studying how an embodied agent can adapt in the real world with noisy perceptions and continuous time and state spaces, and by testing current neuroscience theories in such context; and as a means to test the potential application to robotics by assessing the transferability of neurocomputational models of learning and decision-making to robots operating in dynamic, unknown environments.

This paper is the first to report on spatial navigation with the new version of Psikharpax (v2; figure 1). The robot has been equipped with a rich sensory set of devices for multimodal perception (binaural auditory equipment, artificial whiskers,

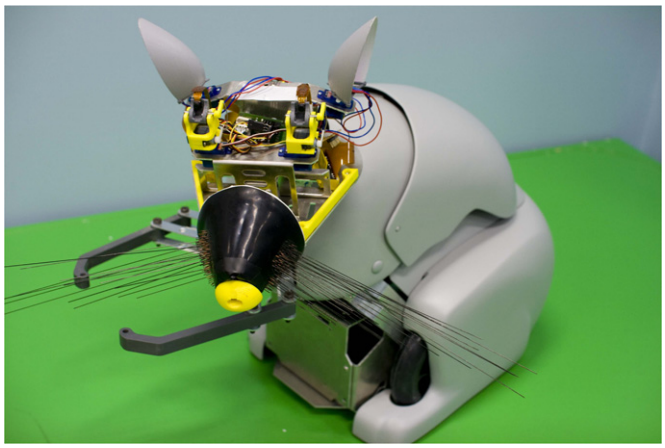


Figure 1. The v2 Psikharpax robot.

binocular vision) and sensory integration. This previously allowed us to perform tactile texture discrimination and obstacle avoidance with the whiskers [2], hearing and noise localization [3], vision and adaptive saccadic eye movements [4, 5]. Here we present the upgrade of the robot's cognitive architecture enabling the robot to coordinate and learn multiple strategies for spatial navigation, and perform fast adaptation to environmental changes.

1.2. Multiple navigation strategies in rodents

Mammals are able to use multiple strategies when faced with a navigation problem [6–9, for reviews], like reaching a hidden platform in a pool, the so-called Morris water maze [10]. Among the numerous possible strategies, experimental neuroscience studies of strategy interactions favored two main families.

- Response strategies, resulting from the learning of direct sensory–motor associations (like swimming toward a cue indicating the platform location, which is called a *taxon strategy*).
- Place strategies, where the animal builds an internal representation (or cognitive map) of the various locations of the environment, using the configuration of multiple allocentric cues. It then uses this information to choose the direction of the next movement either by learning place–action associations (*place recognition triggered response strategy* or *PRTR*) or, more adaptively, by planning a path in a graph connecting the places with the actions allowing the transitions from one place to another (*topological planning strategy*).

It has been shown that the multiple navigation strategies of rodents are operated by parallel independent memory systems [11, 12], which can result in *cooperative* or *competitive* behaviors, depending on the experimental protocol. The basal ganglia (BG) and the hippocampal formation (Hpc) appear to have a central role in this circuitry. The BG can be subdivided into parallel sub-circuits [13], usually identified by the part of the striatum—the main BG input nucleus—they incorporate. The BG operate action selection [14] and use reinforcement learning signals mediated by dopamine [15] to adapt these selections to environmental conditions.

Response strategies are considered to rely on the projections from the sensory and motor cortices to the BG circuits issued from the dorso-lateral striatum (DLS) to select directions of movement, using reinforcement learning capabilities of the BG to learn which cue is to be followed at a given time [16, 17]. Consistently, lesions of the striatum—or more specifically of the DLS—impair or reduce the expression of response strategies while promoting place strategies [18, 19]. In contrast, lesions of the hippocampal system impair place strategies while sparing response strategies [20, 11, 18]. This suggests that response strategies are independent of the Hpc. On the other hand, place strategies would rely on the Hpc, with its ability to encode places in the so-called *place cells* [21], to provide inputs to work with. The neural circuits exploiting them to either learn place–action associations or to plan trajectories would be located in the prefrontal cortex (PFC), in the ventral striatum (VS) and in the dorso-medial BG circuit (DMS). Indeed, lesions of the DMS reduce the expression of place strategies while promoting response strategies [18, 22]. Lesions of the VS impair animals' ability to associate different places with different amounts of reward [23].

1.3. State of the art of neuro-inspired robotic navigation

Several previous projects have tested biomimetic models of rodent navigation on robots, based on these experimental data. Such projects participate in the global approach consisting in transferring neurocomputational models to robotics with a twofold objective: on the one hand, taking inspiration from the computational principles underlying mammals' behavioral flexibility to contribute to the improvement of current robots' autonomy and adaptivity. On the other hand, using the robot as a platform to test the robustness of current biological hypotheses about spatial cognition, beyond perfectly controlled simulations, and try to learn more about the computational mechanisms at stake by analyzing which solutions enabled the model to work on a physical robot [24–26].

Arleo and Gerstner developed a computational model of place cells—neurons located in the hippocampus whose activity encode an estimation of the animal's current position—and head-direction cells—neurons selective for the estimated orientation of the animal's head [27]. With this model, they enabled a Khepera robot to navigate in a small arena, using a navigation strategy where learned associations between places and directions of movement (a PRTR strategy). Fleischer, Krichmar and colleagues showed how prospective and retrospective coding at the level of place cells' activity can enable a robot to efficiently solve a spatial memory task [28, 29]; here also, navigation was performed by a PRTR strategy. Barrera and Weitzenfeld proposed a hybrid PRTR strategy using a graph, where the choice of the next action took into account the next three actions in a prospective manner [30]. Their robot could solve discretized implementations of various rodent laboratory mazes (T and radial mazes). Giovanangeli and Gaussier developed a model of another

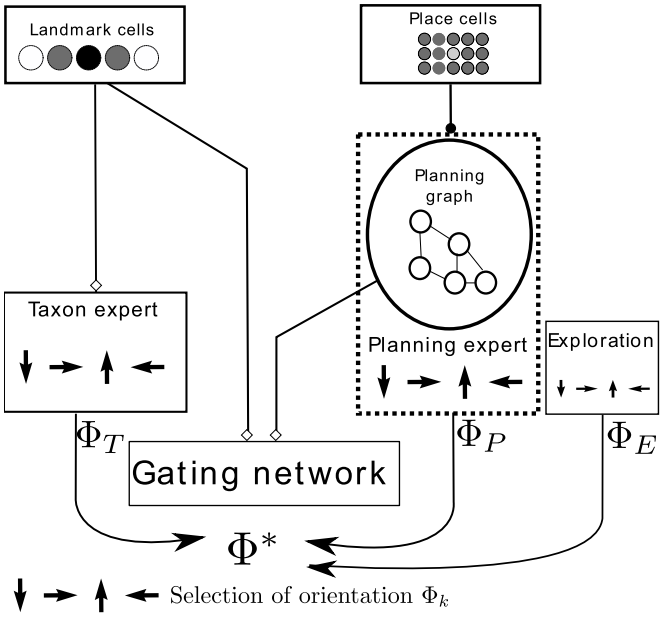


Figure 2. Overview of the Dollé *et al* model [33]. Different strategies (taxon/topological-map/exploration) are connected to the gating network. Each strategy has a dedicated expert which proposes actions (Φ_T for the taxon, Φ_P for the planing, ...). The gating network decides which of the experts is the winner in the current situation and then the action Φ^* from this strategy is performed.

navigation strategy consisting in planning routes toward the goal in a topological graph ('cognitive map') of the environment. Their model produced efficient navigation in both indoor and outdoor environments [31]. More recently, the RatSLAM algorithm has been implemented as a neural network inspired by the rat's hippocampus in order to perform efficient, continuous and long duration simultaneous localization and mapping (SLAM) on a robotic platform put in a large non-stationary environment [32]. Planning is also used here to perform navigation.

Our contribution relies on transferring to robotics another aspect of rodent navigation abilities: the combination of various navigation strategies, in order to benefit from their respective strengths (accuracy, learning rate, adaptation to changes, etc), coordinated by a meta-controller for strategy-shifting which has been previously shown to better reproduce rodents' behavioral performance than single navigation strategies [33]. Thus, we extracted the principles of each previously studied components of rats' currently known cognitive architecture for navigation: place cells, path integration component, path planner, reinforcement learner. And we focused on the integration of these components in a brain-inspired system used for adaptive strategy shifting.

1.4. The computational model previously used in simulation

In this paper, we first apply to the Psikharpax robotic platform the multiple strategy switching model (see figure 2) proposed in [33], which was tested in simulation to replicate rat behavioral experimental results. We then propose an extension of this model allowing a more flexible adaptation when switching from one experimental context to another (i.e. change in goal location).

The model in [33] provides a simple mechanism able to replicate experimental results obtained in [34] and [18] in variations of the Morris water maze protocol: it proposes that a gating network is dedicated to the selection of the strategy to be used, and that it uses reinforcement learning to learn which strategy is the most efficient in each situation, based on all the inputs used by the strategies to take their own decisions (i.e. sensory and place cells activity). All strategies learn simultaneously: those which did not have the control over the last decision use the reward/punishment signals modulated by the angular difference between their movement suggestion and the actual one: the smaller the difference, the more the suggested movement of a non-selected strategy will be rewarded. This is a key element of the model to explain the cooperative effects observed in animals, where the learning process of a slow learning strategy can thus be guided by the selections made by a fast learning one.

In the experiment from [34], external visual cues, allowing the generation of an internal map, are provided, and the hidden platform is indicated by a visual cue standing 20 cm away from the platform in a fixed direction (making taxon strategies more difficult to learn and less efficient than when it is directly above the platform). The platform is moved after every session of four trials, so that rats using a map-based strategy perform poorly at the beginning of a new session, while those favoring a taxon strategy are not much affected. Finally, the experiment is carried out with a group of control rats and another one with rats with hippocampus lesions. The model reproduces the differences in performance and in learning dynamics of both groups: lesioned rats learn across sessions, while control ones also learn within sessions, being less efficient than those lesioned at the beginning and better at the end. The model shows that this can be explained by a competition between strategies in the beginning of a session—each strategy leads to a different place, as the map-based one leads to the previous location—and cooperation in the end—once the new location is known, the taxon provides the global direction to reach the platform in the beginning of the trajectory, but the map-based strategy is more efficient than the taxon to precisely lead to the platform location, rather than to the cue, at the end of the trajectory.

In [18], nine sessions are carried-out with four groups of rats (control, fornix-, DLS- and DMS-lesioned rats), external cues are provided and the platform is either hidden (sessions 3, 6 and 9) or visible. The tenth session is a test where the platform is visible but has been moved. The DMS model was not simulated as the precise modification to be applied to the model was unclear (should the map-based strategy or the gating network be affected? And how?). The fornix- and DLS-lesioned groups were simulated by respectively removing the map-based and the taxon strategy. The main characteristics of the groups' behavior is well captured: when the goal is visible, any strategy can lead to the platform; but when it is hidden, the fornix-lesioned group performs poorly; finally, during the test, the DLS-lesioned group performs poorly as it goes to the previous location, while the control group is not as good as

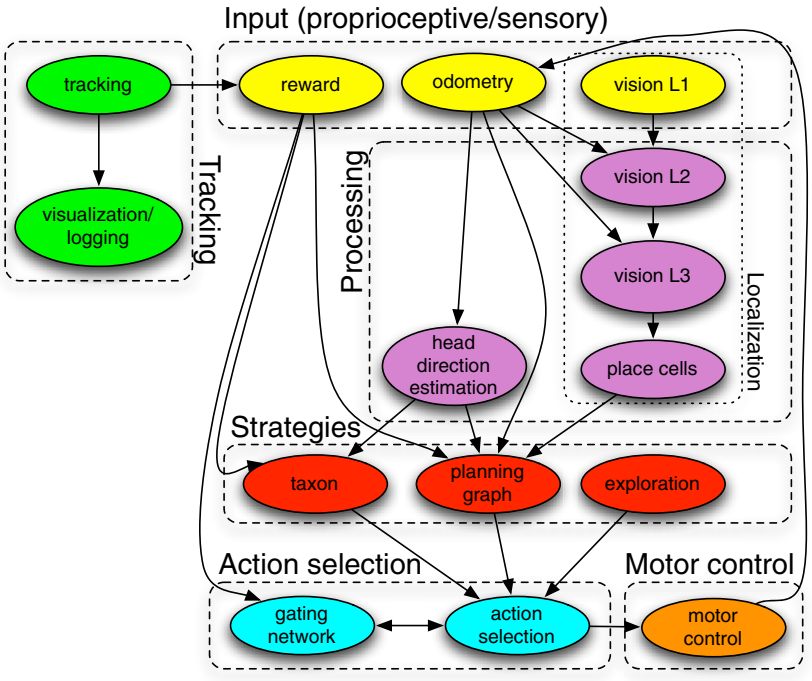


Figure 3. Simplified overview of the software architecture (only the most important nodes and connections are shown).

the DLS-lesioned one, as competition occurs between the two strategies. We refer to the original paper for more details on these two simulations [33].

These results were however obtained in simulations which, although in continuous state space, were perfectly controlled and thus permitted a set of crucial simplifications.

- The model had perfect access to its position and orientation.
- Visual perception was also perfect, permitting the robot to distinguish without errors different landmark cues, and thus making it possible for the model to have taxon submodules which learned to select a movement direction in association with each specific landmark.
- The agent was a virtual point without a body surface, allowing holonomic motion.

Thus, it is not clear whether the model can be applied to robotics in the real world, and whether it can still reproduce rodent behavioral performance and adaptivity in such circumstances.

Here we present the integration of this neurocomputational model in the Psikharpax robot, and the solutions adopted to cope with noisy perception and odometry. While in simulation, each strategy could individually solve the rat goal-seeking task but a combination of strategies was required to produce the same behavioral performance as the rat [33], here we find that each strategy can only partially but complementarily solve the task, and the combination of strategies permits us to achieve the problem. In addition, we show that the previous strategy-shifting mechanism can adapt to environmental changes, but with slower performance than real rats. We finally add a meta-controller to the model which detects context-switching, permits faster adaptation to environmental changes, and allows us to quickly restore previously learned behavior when a known context is presented again to the robot. Such meta-controller may constitute a better model of rat prefrontal cortical functions known to be required for adaptive strategy shifting [35–37]. It may also provide a more robust solution for strategy shifting in autonomous robots.

The first part of this paper gives a technical overview of the platform. The theoretical foundation of our work was verified by Dollé *et al* [33] in simulation, based on almost perfect sensory input and simulated grid cells [38]. Therefore, the second and third parts of this paper present the equivalent navigation strategies and strategy selection mechanism for the real robot. The last part presents the results obtained in a series of robotic tests of the model.

2. Material and methods

2.1. Architecture overview

Our software architecture was built on the ROS[6]—robot operating system—middleware. The robot runs the ROS core and an external quad-core machine is used for the visual system and the navigation strategies.

An overview of the software architecture of our model is given in figure 3. The system consists of six distributed subsystems, each consisting of one or more ROS nodes. As can be seen from figure 3, the central node of the system is the action selection node. This node interacts with the gating network (see section 4) to decide upon the next action the robot will take.

Two additional mechanisms—guiding and obstacle avoidance—are not shown in the figure. The obstacle avoidance strategy is implemented as a reflex strategy to

[6] ROS is an open-source system and can be downloaded from http://ros.org.

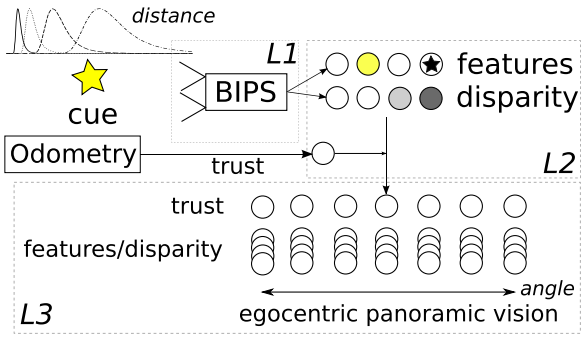


Figure 4. Concise overview of the visual system. In this example, there is a brightly colored star-shaped object at a distance of approx. 3.5 m from the robot's head. The robot sees this object through its two cameras directly connected to the BIPS (bio-inspired perception system) hardware (L1). The BIPS hardware extracts feature information from the visual object and this information is coded on a set of feature neurons in the second layer of the visual system (L2). Based on the angle at which both cameras see the cue, the disparity neurons are activated to code the distance information (see also figure 6). The trust neurons are activated based on odometric information: if the robot's head is moving fast, the trust drops. There are disparity, feature and trust neurons for each direction within the field of view (not shown in the figure). Information in the L2 layer is sent to layer L3 and integrated over different orientations to produce a 360 degree view.

prevent the robot from leaving the environment. The guiding procedure is used to lead the robot toward and from the goal at the end of failed and successful trials (see section 5.4).

2.2. Visual processing and localization

More details on the visual system are given in appendix A of this paper. Here we summarize how visual information concerning landmark cues in the environment is extracted to build a map of place cells for localization. An overview of the model is given in figure 4.

The robot is equipped with two small front-facing cameras with a total field of view of about 60 degrees. While real rats have side-facing eyes with a large field of view, their stereoscopic vision is limited to a region of about 76 degrees [39]. The choice for a small but stereoscopic view originally stems from experiments with saccadic eye movements on the Psikharpax platform [5]. We chose to keep this setup as it allows the robot to estimate the distance of objects and it allows us to extend the model to include attention. To overcome the limited field of view, the robot is programmed to turn its head around at regular intervals.

At the lowest layer of the visual system, the cameras are directly connected to an onboard electronic device, called the bio-inspired perception system (BIPS, developed by Brain Vision Systems, BVS) [4]. This layer implements the retina and the first layers of the visual cortex using a neural network with inhibitory connections to detect and track stable and saturated objects (see figure 5). This layer is shown as L1 in figure 4. Note that after this layer, the raw image is discarded and only the detected objects are used.

The second layer of the visual system codes the visual information onto a set of feature neurons. For each object, the following features are extracted: size, vertical position, orientation, color and disparity. For each orientation within the field of view, such a set of feature neurons exists and a detected object activates the neurons in the direction in which it is seen. We use leaky-integrator neurons to low-pass filter the input. The disparity codes the distance of an object with respect to the robot. Four neurons are used to code disparity information. These neurons have a Gaussian activation function, centered around different disparities. This results in an activation function that has a large tail as a function of the distance (figure 6). Hence the robot has more precise distance information on nearby objects.

Primates seem to use other types of disparity measures such as relative disparity between objects next to absolute disparity [40]. We tried to increase the performance of the visual system by adding disparity neurons with other activation functions (based on Gabor filters), but the quality of the resulting place cells (see the following section) did not increase. This is probably due to the fact that enough information to distinguish places can already be extracted by the nonlinear training algorithm from the other feature neurons.

An important part of the second layer of the visual system are the so-called trust neurons. These neurons modulate the

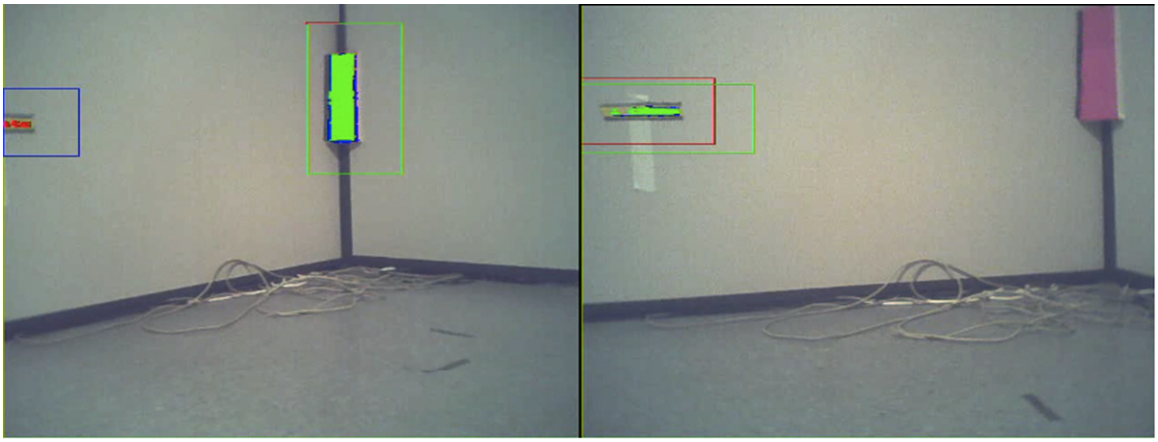


Figure 5. The same part of the environment seen from two different angles can result in an object not being detected (object on the right). The system should cope with this limitation through the higher layers of the visual system.

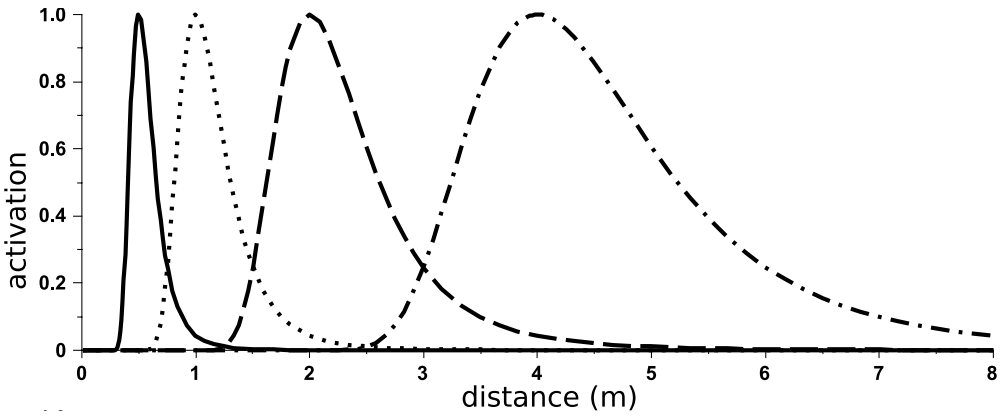


Figure 6. Activation functions of the disparity neurons as a function of the distance. They are Gaussian as a function of the disparity, naturally resulting in a nonlinear distance scale with higher precision for nearby objects.

output of the second layer to the third layer of the visual system. The idea is to suppress noisy inputs when the visual input is unreliable. This occurs when the head of the robot moves too fast, as the neurons in the first (tracking units) and the second layer need some time to stabilize. This is easily detected by the odometric system and hence the odometric system is used to modulate (suppress) the connections between the second and third layers when necessary. The faster the head movement, the less reliable the visual information. This prevents the third layer of the visual system from being influenced by unreliable information.

The third level of the visual system integrates the information from the second layer by combining it with odometric information. This results in egocentric panoramic information on the environment.

2.3. Visual place cells

The output from the neural network visual layers is high dimensional (about 800 neurons). Because simple rate-coding neurons were used, the output can be seen as a vector representing egocentric visual information integration. To construct non-directional place cells, such output vectors were summed over all orientations to activate the same neuron (i.e. a place cell). The problem was therefore reduced to a dimensionality reduction or clustering problem. This subsystem is indicated as *PC* on figure A1 in appendix A.

In a first version of the simulation model [33], ad hoc place cells were used, and thus the dimensionality reduction/clustering problem was not addressed. In [41], a model of the hippocampus [38] was used to autonomously create the place cells. It is based on a competitive Hebbian-like learning rule: a number of random place cells are created; during the learning phase, the place cells specialize for particular input patterns using a sparseness-based Hebbian rule, which only allows for the most active input neurons to reinforce their connections.

Such an approach works very well when the number of input neurons and distinct patterns is not too high and the patterns are well characterized by their most active neurons. In our case, however, the input can be noisy with typically large but meaningless values for a few neurons in the input. When the sparseness function from [41] is applied to such an input, the noise is reinforced, while useful neurons are ignored.

We therefore needed a technique that learns the input patterns by evaluating the whole set of input neurons instead of only the most active ones. We initially tried linear approaches such as principal component analysis [42] to check if the inputs were linearly separable. At most four to five regions could be consistently separated. This is insufficient for good performance, as the place fields of the place cells would be too large (only four to five distinct zones). Indeed, the gating network takes input from the place cells and hence its precision is limited by the place cells.

2.3.1. Implementing a SOM. A popular nonlinear alternative for clustering consists in using self-organizing maps (SOM) [43]. The goal of SOM is to move the neurons in the high-dimensional input space to approach the topology of the input. For each input, the Euclidean distance between the input and each neuron of the SOM is computed. The closest neuron is called the best-matching unit (BMU). The SOM is then updated by moving the BMU closer to the input (weighted sum) as well as its neighbors. In order to avoid using a fixed number of neurons in the SOM, we used the growing neural gas (GNG) algorithm [44]. GNGs are created incrementally by inserting a new neuron after a number of input samples by splitting the neuron with the largest accumulated error (sum of distances) into two new neurons. The topology itself is also learned by keeping the neighborhood of the neurons up to date.

We used the GNG algorithm to learn the weights of an artificial neural network (see appendix A.2 for a detailed description of the implementation). While we do not assume that there is a direct biological equivalent of this training algorithm, nor the activation function (which is based on the Euclidean distance), we do not think that our model makes unrealistic assumptions about the role of the hippocampus in categorizing different places. As [44] indicates, the GNG algorithm can be seen as a form of (nonlinear) competitive Hebbian learning, which is the main reason why we chose this algorithm. Because the main interest of this work lies in the strategy selection and context-switching mechanisms, we did not investigate how exactly one might implement the

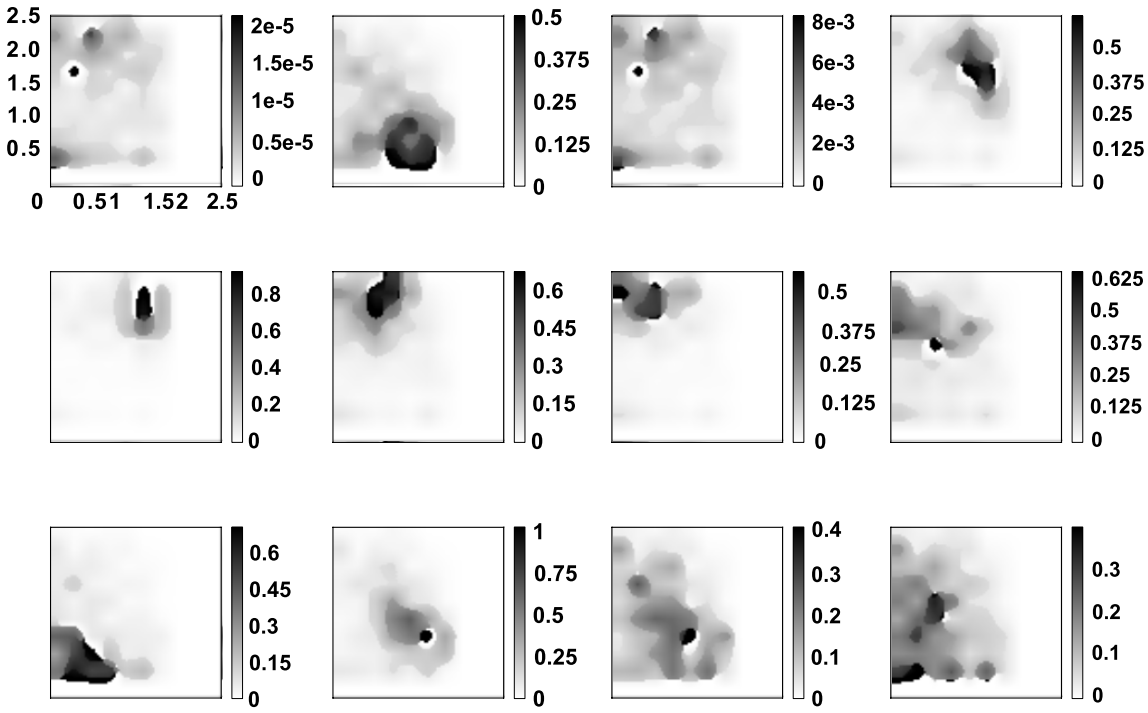


Figure 7. Heat map of 12 place cells (GNG with 12 output neurons), with smooth activation (equation (A.4)). Note that there are two place cells (top left and first row third from left) which only have very weak activations but large receptive fields. These cells will thus not be important as their activation will be negligible (the topological map will discard them). The axes of each of the images give the position in meters of the robot's head as recorded by a ceiling camera. Note that there are more and more precise place cells coding for locations near the borders of the environment.

GNG algorithm with a biologically plausible neural network. However, this should not be a problem, as the algorithm is straightforward and only depends on the computation of the Euclidean distance (or another distance measure), the creation/removal of edges and updating a local error measure. A further argument for using a nonlinear approach is that nonlinear algorithms can often be cast as a linear technique working in a larger (possibly infinite dimensional) feature space by simply replacing inner products with a kernel function (i.e. the kernel function computes the inner product in a different space) [45]. For example, the kernel trick is often applied to PCA (Kernel PCA) [46] for which a Hebbian version already exists [47].

We found that the GNG technique is very flexible. When one forces the GNG to use only a small number of neurons, the GNG creates large continuous place fields (figure 7). When more neurons are available, the place fields are smaller.

The neurons from the GNG layer project onto the planning strategy and the gating network. The resulting place cells for a GNG layer with 12 neurons are shown in figure 7. One can see that the best (i.e. most restricted to a particular zone) place cells are found near the borders (similar to figure 9(D) of [9]). This is due to the fact that there are no intramaze cues, which causes observations near the center to be more similar and thus less place cells are created in this region. We used a higher number of place cells in our experiments to increase the precision of the planning strategy (section 3.1) and the gating network (section 4.2). However, this phenomenon still occurs (e.g. figure 8(*b*)), causing topological maps to be less dense near the center.

2.3.2. Experimental testing of the place cell system. To get a rough estimate of the usefulness of the place cell system, the robot was put at 40 random places in the environment, where the activation of its place cells were recorded (for this experiment, we trained 100 place cells). The Euclidean distance between the real position of the robot and the center of activation (computed by averaging over a large training set using a ceiling camera) of the most active place cell (binary activation) was computed to estimate the precision of the place cell coding. The mean distance was found to be 16.5 cm with a standard deviation of 8.5 cm. When near the border of the environment, the mean lies around 11 cm, which is very good, but when approaching the center of the environment the mean distance becomes considerably larger (between 20 and 30 cm). In [9], the authors found a mean error of 6 cm but in a smaller (0.8 m by 0.8 m) environment.

In the ideal case, one would expect the place fields to be evenly distributed. Hence, to evaluate our place cell mechanism, we consider 100 points picked randomly from a uniform distribution on a rectangle of 2.5 m $\times$ 2.0 m. The expected distance from any such point within the rectangle to the nearest point out of the 100 randomly chosen points is about 12 cm. The expected distance to the second closest point is about 18 cm and 22.5 cm for the third closest point.

This indicates that the presented place cell system performs reasonably well, compared to the ideal, uniformly distributed case.

We tested the system with different numbers of place cells, by splitting the data into a train and test set. We found

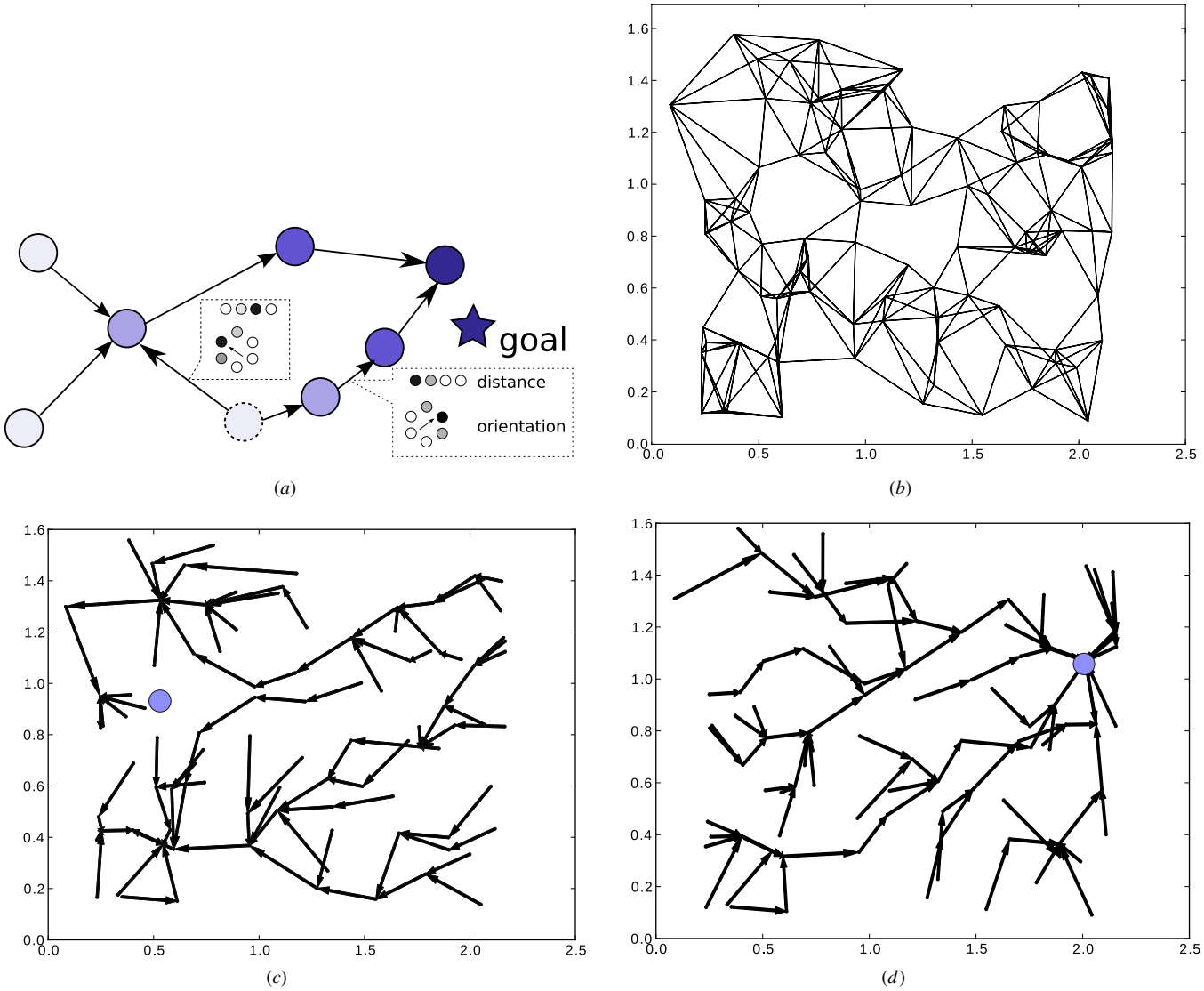


Figure 8. Overview of the topological map. The map is the same in (*b*), (*c*) and (*d*). (*a*) Principle of operation. (*b*) Constructed topological map after exploration. (*c*) and (*d*) Path planning using the topological map. The qualitative difference between (*c*) and (*d*) is caused by the change of the goal location. In (*c*), there is a lot of aliasing around the goal location (multiple high diffusion values), resulting in two disjunct path-planning trees and sub-optimal paths near the goal (e.g. there is an optimum around (0.5, 0.4)). In (*d*), the topological map works very well (one tree with almost every edge leading the robot closer to the goal). Note also the difference in map quality between the simulation model (e.g. figures 8(*d*) and 12(*b*) in [33]) and the robotic platform. This is caused by the extensive exploration phase in simulation and the noise-free simulation environment. (*a*) Principle of the topological map. The large circles are the nodes in the map. The goal is on the right and the color of the nodes correspond to the diffusion values. Direction transition neurons (small circular pattern) and distance transition neurons (four small circles) are shown for two connections. A path planned from the dashed node can be ambiguous for one node due to equal diffusion values. The robot chooses a next node arbitrarily in this case. (*b*) Topological map constructed by the robot. Vertices correspond to nodes in the map, and edges are paths the robot can use. The maximum number of neighbors per node was fixed at 6. (*c*) Paths toward the goal shown from each node in the topological map. In this map, there is aliasing around the goal, resulting in very low performance of the planning expert on the left side of the environment, near the goal (circle). Axes are in meters. (*d*) Paths toward the goal shown from each node in the topological map. This map predicts good performance of the planning expert with only a few regions with ambiguities.

that 100 place cells is about the maximum one can obtain in our environment without overfitting the training data. Moreover, for higher numbers of cells, the classification results on the test set did not increase. Thus, for all the next experiments with the multiple navigation strategy model, the number of place cells was fixed at 100. In the next section, we present the navigation strategy which is based on information from the place cell layer.

3. Navigation strategies

3.1. Planning expert

The model in [33] best replicated experimental results using a planning algorithm for the place strategy, rather than a place recognition triggered response one. It is organized as follows: a place cell module, simulating the hippocampus, is

in charge of learning internal representations of places in the environment using sensory inputs; a graph module (topological map), simulating the PFC, learns by means of a Hebbian rule the directions of movement, which are used to go from one place to another. When the goal has been found at least once, a diffusion of activity in this graph originating from the goal node generates a gradient which, when followed, leads to the goal with the shortest path [48, 49].

The simulation model from [33] uses distinct representations for place cells and nodes in the graph module, because they differ in function and precision. More precisely, a simple single-layer network trained with a competitive Hebbian-like learning rule is used to activate the topological map nodes based on the place cells' activation. Because of this layer, the number of topological map nodes (around 100) was typically a factor 10–20 lower than the number of place cells.

Because of the encouraging results from this simulation model, we chose to adapt it to the physical platform. Several modifications needed to be made to make this feasible, which will be explained in this and the next sections.

The maximum (useful) number of place cells on the physical robot is limited by the quality of the sensory input (see the previous section). Because the goal is relatively small and a high level of detail in both the planning strategy and the gating network is advisable to get precise results, we mapped the place cells directly onto the nodes in the topological map to get the maximum resolution. This means that there is no additional training to map the place cells onto the topological map nodes (1-to-1 connections).

We initially also tested (not shown) the system with a coarser representation of space (using an additional layer trained with competitive Hebbian learning) for the gating network, yielding similar, but less detailed results than the ones presented in this paper. The added benefit of the direct mapping from place cells to nodes in the topological map is that analyzing the results is easier as the space representation is the same throughout the system (gating network, place cells and topological map). To underline the functional difference between nodes in the topological map and place cells, we use the notation n^{PFC} for nodes in the map (referring to the PFC) and n^{PC} for place cells.

3.1.1. Learning the topological map. During an exploration phase, the topological map learns connections between nodes by Hebbian learning. For this, two types of information need to be stored, the relative angle between two nodes and their mutual distance (figure 8(*a*)). To store this information, we use two sets of transition neurons for each connection between nodes [50, 51]. There are N_{ANG} transition neurons (per set) for directional information and N_{DIST} for distance information. Each node initially has connections (transition neurons) to every other node with zero weights. The transition neurons are stored in a vector $v_{k,l} = [v_{k,l}^1, \ldots, v_{k,l}^{N_{\mathrm{ANG}}}, v_{k,l}^{N_{\mathrm{ANG}}+1}, \ldots, v_{k,l}^{N_{\mathrm{DIST}}+N_{\mathrm{ANG}}}]^T$, i.e. the subscript k, l indicates the transition from node k to l and the superscript is the affected transition neuron (orientation or distance information).

As for the distance information, the angular information is stored in a set of N_{ANG} neurons with Gaussian activation functions centered around a fixed directions. We define the vector $b = [b^1, \ldots, b^{N_{\mathrm{ANG}}}, b^{N_{\mathrm{ANG}}+1}, \ldots, b^{N_{\mathrm{DIST}}+N_{\mathrm{ANG}}}]^T$ similar to $v_{k,l}$. The first N_{ANG} elements of b code the angular information between locations. The last N_{DIST} neurons contain the distance information between two locations. That is, a vector $b(t_0, t_1)$ contains the activation of the transition neurons to the location where the robot is at time t_1 from the location where the robot was at time t_0.

Now, to update the neurons $v_{k,l}^i$, we iterate over a trajectory of the robot and update the weights to the transition neurons using a simple learning rule:

$$\Delta v_{k,l}^i(t+1) = \begin{cases} (1-\delta_{k,l})H\left(n_{\mathrm{conf}}^{\bar{L3}} - \beta\right)b^i(t_k, t) & \text{if } k = \text{last} \wedge \text{l} = \text{winner} \\ 0 & \text{else.} \end{cases}$$

Here $\delta_{k,l}$ is the Kronecker delta, to prohibit connections from a node to itself. $H(x)$ is the Heaviside step function used to prevent updating the graph when the confidence of the place cells $\left(n_{\mathrm{conf}}^{\bar{L3}}\right)$ is below a threshold β. *last* is an index referring to the node in the topological map where the robot was when $n_{\mathrm{conf}}^{\bar{L3}}$ was above the threshold for the last time, i.e. the previous location.

For planning in this graph, only the shortest transitions between nodes are kept, based on the distance coding neurons. For the results presented here, we fixed the maximum number of transitions per node to 6. A learnt map is shown in figure 8(*b*). While we explained this process as a sequential algorithm (recruitment of nodes, learning of transitions, competition), it can be done online by simply adding an additional transition usage intensity neuron $v_{k,l}^0$ increasing in activation when the robot moves from k to l, combined with a decay rate or competition factor (i.e. $-\psi v_{k,l}^0$) to the previous equation, which also prevents the weights from increasing without bounds.

3.1.2. Using the planning expert. In order to plan in this graph, the model maintains a set of neurons g_i, one for each node corresponding to the reward received at each of the locations. A leak rate is added so that the robot can navigate in an environment with changing reward locations. g_{winner} is the neuron assigned to the node at the robot's current position.

$$g_j(t+1) = g_j(t)(1 - \tau_{\mathrm{forget}}) \tag{1}$$

$$g_{\mathrm{winner}}(t+1) = g_{\mathrm{winner}}(t)\left(1 - \tau_{\mathrm{learn}} n_{\mathrm{winner}}^{\mathrm{PFC}}(t)\right) + \tau_{\mathrm{learn}} n_{\mathrm{winner}}^{\mathrm{PFC}}(t) R(t). \tag{2}$$

The second value associated with a node is the diffusion value $d_j(t)$ and this value is used to implement a shortest-path algorithm. The activation from the goal diffuses or spreads out [48, 49] over the other nodes (figure 8(*a*)). To compute the equilibrium state efficiently, we used a modified Floyd–Warshall algorithm [52], where $d_j[iter]$ is used to refer to the value of d_j at iteration *iter*:

Algorithm 1 Computing the diffusion values

$iter = 0$
for $i = 0; i < N_{\text{PFC}}; i = i + 1$ **do**
  $d_i[0] = g_i$
end for
while $iter < N_{\text{PFC}} - 1$ **do**
  for $i = 0; i < N_{\text{PFC}}; i = i + 1$ **do**
    $d_i[iter + 1] = \max(d_i[iter], \max_{j \in neighbors(i)}(d_j[iter])\iota)$
  end for
  $iter = iter + 1$
end while
for $i = 0; i < N_{\text{PFC}}; i = i + 1$ **do**
  $d_i = d_i[N_{\text{PFC}} - 1]$
end for

This algorithm is only run when $n^{\bar{L3}}_{\text{conf}} > \beta$, i.e. when the robot has a high trust in its current position.

The algorithm finds the shortest path to a maximum goal value (g_j) in terms of the number of intermediate nodes and the goal values. However, multiple maxima can exist as more than one node can code for the goal location due to aliasing. To find a path to the closest maximum (goal), one starts from the current node ($n^{\text{PFC}}_{\text{winner}}$) and chooses the neighbor with the highest value as the next node on the path. The topological map then proposes the mean of angles $\Phi^P(t)$ computed from the activation of the direction transition neurons associated with the connection from $n^{\text{PFC}}_{\text{winner}}$ to its most active neighbor.

In practice, we add a slight twist by storing a complete path toward the closest maximum. Next the robot computes the relative positions of the nodes of the path in the egocentric frame. It then tries to follow this path by moving sequentially to the position of the nodes based on the odometry. This allows this map-based strategy to persist even in the absence of sensory input and is equivalent to the basic algorithm when sensory input is reliable.

Figures 8(*c*) and (*d*) show the same map with two different goal locations; the arrows represent the direction toward the goal from each node in the graph, as computed by the planning algorithm. In the situation shown in figure 8(*c*), the robot faces an aliasing problem: there are multiple optima (high diffusion values) near the goal, and starting from the east part of the environment, the robot may plan trajectories toward different nodes apparently close to the goal. As a consequence and as we will illustrate in the experiments with the whole model, the planning expert can adapt quickly but remains approximative. It can quickly learn trajectories toward the coarse area around a new goal location. But these trajectories may not be precise enough to reach the goal and other experts using different strategies may be more relevant in more precisely attaining the goal location.

3.2. Taxon strategy

The second strategy is implemented in the so-called taxon expert. It learns to associate proximal visual cues with actions using a standard Q-learning algorithm [53]. While in previous work with noisy continuous state space in complex mazes we employed a multiple-module reinforcement learning approach for the taxon system [54], here we used a simplified taxon in order to test the ability of our meta-controller to switch between complementary strategies. While rodents' hippocampus-dependent place strategies rely exclusively on distal landmarks—which are far and outside the maze and thus are more stable to constitute the anchoring of a cognitive map—the taxon strategy consists in learning directions of movements in association with intra-maze proximal landmarks [55, 18, 56].

Thus here the taxon expert can only perceive the goal. In order to prevent the robot from seing the goal when it is far—thus distal—we make its perception noisy. Thus, the relative position of the goal is seen by the taxon as a Gaussian which decreases in height and variance as the distance increases (until it drops below a threshold). While conceptually simple, it is important to note that the taxon expert initially does not know that it should move toward the stimulus it receives. So the taxon (like any other learning expert) learns at the same time as the gating network.

The update equations of the taxon expert are based on [33], with slight modifications to adapt the strategy to the robot platform.

The possible directions (continuous) the robot can take are coded on the N_{dir} action cells a_i. $w_{i,j}$ are the Q-values, which are used to associate input orientations with output orientations:

$$a_i(t) = \sum_{j=1}^{N_{GP}} r_j^{GP}(t) w_{i,j}(t). \tag{3}$$

By taking the mean of angles Φ^T, we obtain the proposed action of the taxon:

$$\Phi^T(t) = \arctan\left(\frac{\sum_i a_i(t)\sin(2\pi i/N_{\text{dir}})}{\sum_i a_i(t)\cos(2\pi i/N_{\text{dir}})}\right). \tag{4}$$

$\Delta w_{i,j}$ is the update rule for the Q-values, based on the reward-prediction error δ^{taxon} and the eligibility traces $e^{\text{taxon}}_{i,j}$:

$$\Delta w_{i,j}(t) = \eta \delta^{\text{taxon}}(t) e^{\text{taxon}}_{i,j}(t). \tag{5}$$

The eligibility traces are used to speed up learning by storing previous state-action pairs and by using action generalization. The action generalization is given by r_i^{AC} and is based on the *executed* action, which can be the action proposed by a different strategy (see section 4.2)[7]. Thus, the taxon strategy also learns if another strategy performs well. Note that in practice one uses a wrapped Gaussian as the difference between circular values has to be computed:

$$\delta^{\text{taxon}}(t) = R(t+1) + \alpha \max_i a_i(t+1) - a(t) \tag{6}$$

$$e^{\text{taxon}}_{i,j}(t+1) = r_j^{GP}(t) r_i^{AC}(t) + \kappa e^{\text{taxon}}_{i,j}(t) \tag{7}$$

$$r_i^{AC}(t) = \exp\left(\frac{-(\Phi^*(t) - 2\pi \text{i}/N_{\text{dir}})^2}{2\omega^2}\right). \tag{8}$$

[7] Note that there is a small error in equation (7) of [33]. It is indeed $\Phi^*(t)$ instead of $\Phi^{\text{T}}(t)$.

The input of the system is the N_{GP} goal direction neurons r_j^{GP}, which replace the landmark cells r_j^{LC} from [33]. r_j^{GP} are again samples from a wrapped Gaussian at equally spaced angles. The width of the Gaussian increases as the robot approaches the goal as does its amplitude. The amplitude of the Gaussian drops below the threshold between 0.5 and 0.75 m.

Because the goal is the only landmark that can be seen by the taxon expert, the taxon is essentially the same if the input orientations are egocentric or allocentric. In this work, we used an allocentric taxon because the robot approaches the goal with its head pointing in the direction of the goal. Hence learning the correct action to take when the goal is behind the robot is slower.

3.3. Exploration strategy

This is a simple strategy that proposes random directions and which serves two purposes in our model. We do not however consider this strategy equivalent to the much more complex exploratory behavior in rodents. It has been shown that rats, in a new environment, observe recurrent patterns such as spending the first minutes establishing a home base [57], then moving out slowly in a zig-zag way, and coming back home in a straight line. We just use a simple random exploration strategy in order to make sure the robot would cover most regions of the environment. Because each action proposed by this strategy persists for five steps, it allows for some exploration and to break out of looping behavior. Secondly, this strategy allows us to evaluate the performance of an expert with respect to the chance level. Because its behavior is random, we expect every other strategy to perform at least as well in most regions of the environment. As we will see, this is not generally valid, as a random strategy might outperform a more elaborate strategy in certain situations.

4. Strategy selection

4.1. General framework

The strategy selection model of [33] is based on the premises that rodents have multiple navigation strategies at their disposition to reach a goal and that they are capable of switching between them. These strategies coexist and are learned in parallel and independently, while a strategy selection mechanism learns to associate perceptions (or situations) with a preferred strategy by means of a Q-learning algorithm.

Accumulating evidence support the hypothesis that Q-learning is a plausible mechanism by which part of mammals brain learn by reinforcement [58]. Neural correlates of action values (similar to Q-values) have been found in the BG [59] and correlates of action-dependent reward prediction errors (consistent with Q-learning) have been found in dopaminergic neurons [60] as well as in the medial PFC [61].

A global overview of the system is shown in figure 9. The basic idea is that we have a number of strategies or experts providing the next action to take (following the given strategy) at each time step, while the action selection network selects

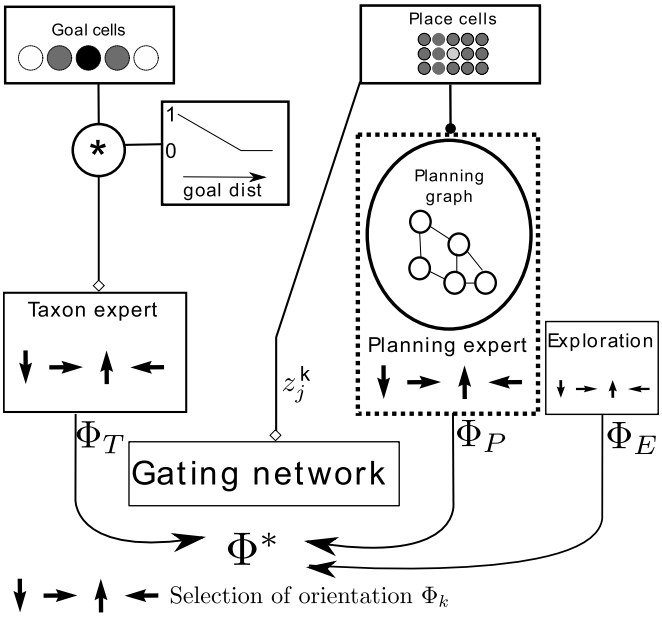


Figure 9. Overview of the gating network with taxon, topological map and exploration experts.

one of these actions, based on the current situation. A Q-learning algorithm based on the simulation model from Dollé *et al* [33] is implemented to allow the robot to associate a state with an optimal strategy. The so-called landmark cells of the simulation model (figure 2) necessitate a mechanism to identify and track distal landmarks in the environment, while here, due to perceptual aliasing and noise in physical experiments with the robot, we used the global configuration of distal cues to build place cells without requiring the recognition of individual cues (2.2). Thus, the connection of the visual system to the gating network is left as future work (see section 7) and in our current model we only use the place cells as input to the gating network. This has the benefit of allowing us to visualize the relationship between locations and preferred strategies as the Q-matrix associates place cells with experts.

The actions proposed by the strategies are the *common currency* in the model as they are the generic information used to evaluate the performance of each strategy [33]. In our model, these actions are the next (egocentric) direction to follow. This way, the robot selects a new orientation provided by the strategy that was deemed the winner among the strategies, turns and moves forward for a fixed distance, observes its new situation (state) and the obtained reward (if any), and finally updates its strategy selection network.

The neural network used to implement the strategy selection is a single-layer network, with the neurons coding for the current situation or perception fully connected to each of the output neurons, one for each available strategy. The layer of output neurons with the connections to the input neurons is called the *gating network*, as it only lets one action through at each time step. One can of course consider the winner of a lower-level gating network as the input of the current network to create a hierarchical selection mechanism.

4.2. Gating network

The gating network computes the so-called gating values $g^k(t)$, one for each strategy k. The Q-values are stored in a matrix $z_j^k(t)$, associating inputs from the place cells with gating values:

$$g^k(t) = \sum_j^{N_{PC}} z_j^k(t) n_j^{PC}(t). \tag{9}$$

Instead of adopting the winner-takes-all policy ($\Phi^*(t) = \Phi^{\text{argmax}_k(g^k(t))}(t)$) from the simulation model to select the winning strategy for the next action, we generalize such a principle so that the selection probability of an expert increases with its relative gating value

$$\mathrm{P}(\Phi^*(t) = \Phi^k(t)) = \frac{g^k(t)^\zeta}{\sum_i g^i(t)^\zeta}. \tag{10}$$

Here $\Phi^k(t)$ is the action proposed by expert k at time t. $\Phi^*(t)$ is the final action proposed by the gating network. Note that this action is not always the executed action, as higher priority mechanisms can override the gating network (i.e. obstacle avoidance or guiding). For $\zeta = \infty$, our action selection mechanism is equivalent to the one from [33]. For $\zeta = 1$, one obtains the action selection mechanism from [62]. For all experiments in this paper, we set $\zeta = 1$, except for figures 30(*a*) and (*b*) for which $\zeta = \infty$ to show the behavior of the robot when it follows its learned optimal policy.

The advantage of introducing some randomness in the action selection is that slower learning strategies can catch up with fast learning strategies when they start to perform better only after a long time. With a winner-takes-all strategy, one might have to wait for convergence before a slower learning but optimal strategy can increase its weights beyond those of a faster learning but suboptimal strategy. This is also biologically relevant since choosing a suboptimal strategy from time to time allows for exploration of unfamiliar alternatives [63].

Learning is sped up using action generalization and eligibility traces. The equations for these techniques were taken from [33]. However, a substantial difference lies in the equation to update the eligibility traces. Whereas sensory input is always reliable in the simulation model, it is not true in general with the real robot. To incorporate this fact in our system, the eligibility traces are modulated by the trust neurons introduced in section 3.1:

$$e_j^k(t+1) = n_{\text{conf}}^{\bar{L}3}(t)\Psi(\Phi^*(t) - \Phi^k(t))r_j^{PC}(t) + \lambda e_j^k(t). \tag{11}$$

To update the Q-values, a modified Q-learning algorithm [64, 53] is applied:

$$\Delta z_j^k(t) = \xi\delta(t)e_j^k(t+1) \tag{12}$$

$$\delta(t) = R(t+1) + \gamma \max_k(g^k(t+1)) - g^{k^*}(t) \tag{13}$$

$$e_j^k(t+1) = \Psi(\Phi^*(t) - \Phi^k(t))r_j^{PC}(t) + \lambda e_j^k(t), \tag{14}$$

where ξ is the learning rate of the algorithm and δ the reward prediction error.

The reward prediction error δ is based on the observed reward when performing action $\Phi*$ and the future expected reward (g^{k^*} is the activation of the winning output neuron and γ the future reward discount factor). The eligibility trace e_j^k reinforces previously selected strategies and the strategies proposing a direction close to the one proposed by the winning strategy [53]. Here λ is the decay factor for previous winning strategies and Ψ a Gaussian function:

$$\Psi(x) = \exp(-x^2) - \exp\left(-\frac{\pi}{2}\right). \tag{15}$$

The gating network is a simple but effective way to combine competition and cooperation between strategies. While the gating network itself only directly provides competition, the strategies cooperate by sharing rewards and their actions (e.g. the taxon uses the executed action for learning, instead of its proposed action (equation (8))). Hence, the gating network is advantageous for strategies as they can learn from each other, while at the global level the performance can also increase because the best performing strategy can be used in each situation.

4.3. Context-switching meta-controller

The advantage of using a gating network for strategy shifting is the possibility of memorizing that strategy A is efficient in a subpart X of the environment while strategy B is relevant in subpart Y [62, 33]. However, since this is learned through Q-learning, noisy information perceived by a physical robot may render this process very slow. In addition, a change in the environment requires us to unlearn A–X and B–Y associations (which can also be very slow). As a consequence, these associations cannot be used again if the environment comes back to its previous state (i.e. the animat cannot recall what it has previously learned).

To overcome this limitation, we implemented a simple context-switching mechanism. The idea is that a change in the environment (e.g. a change in goal location) will be quickly reflected in the profile of diffusion of goal information in the topological map once the robot found the new goal. Such profile can be identified as a context, and the system can recognize a previously experienced context when the goal is set back to its initial position and the model diffuses such goal location in the topological map. Each time the model detects a new context, it will create a new memory component to store values of the gating network in the new context, without erasing values of the gating network associated with the previous context. This part of the model may be viewed as a primitive PFC-based cognitive control mechanism allowing us to associate different contexts with different task sets [65] (see section 7 for a detailed discussion).

In practice, before every step, the gating network decides upon the context it is working in. For this it uses the current vector of diffusion values d from the planning strategy. The gating network now stores a set of Q-value matrices $z_{i,j}^i$, and associates a diffusion vector u^i with each matrix. The current context is now chosen as follows (d is the current diffusion vector):

$$v^i = \frac{d \cdot u^i}{\|d\|\|u^i\|} \tag{16}$$

$$z_{k,j}^* = z_{k,j}^{\text{argmax}_i v^i}. \tag{17}$$

When $\max_i v^i$ is below a threshold, a new context is recruited.

5. Experimental setup

5.1. Introduction

The robot is allowed to move in an open 2 m × 2.5 m environment (figure 18). There are only extra-maze cues (we tested with 10, 13 and 18 cues) and the position of the robot is defined by the position of its neck. The egocentric reference frame has the neck of the robot as its origin and the orientation is defined by the direction of the head.

5.2. Action selection

The robot makes discrete movements, moving 10 cm at each time step. The action selection mechanism is a simple finite state machine that waits for all strategies to propose an action (an egocentric direction) and then activates the gating network to find the winning strategy. The action proposed by the winning strategy is then executed, except if a higher than normal priority mechanism proposes an action (guiding/obstacle avoidance). After an action is performed, the reward is used by the gating network and all strategies to update their learning parameters.

5.3. Reward

The reward node is a simple node processing information from the ceiling camera. When the robot's head passes through the zone defined as the goal, the reward is 1, else the reward is 0. In all experiments, the goal diameter was set to 20 cm (314 cm^2 or 1/160th of the environment).

This is a global reward signal, shared by all strategies and the gating network. Strategies learn by updating their parameters using the global reward.

5.4. Experiments

All experiments consist of two phases. During the first phase the robot explores the environment. In this phase, we force the robot to visit a number of locations in the environment so that enough information is available to construct place cells and a topological map (but no diffusion values nor goal values). The gating network and the navigation strategies are not active and there is no reward in the environment. This phase can be common to multiple experiments. During the second phase, the robot is put at a random location in the environment and the gating network and navigation strategies are turned on. In this phase, the strategies and the gating network use the place cells and topological map to learn to navigate toward a goal.

A goal location is chosen and the robot learns to appropriately coordinate navigation strategies to reach it. During an experiment, after each failed trial (i.e. the robot does not find the goal after 5 m of movement), the guiding procedure forces the robot toward the goal to show the goal location and thus speed up learning, similar to the procedure used by experimenters in rodent laboratory tasks. This is achieved by using the ceiling camera that tracks the robot's position and provides actions leading toward the goal in a straight line. The robot learns the result of this movement as if it had decided itself to perform it. Similarly, when the robot has received a reward (i.e. a successful trial), the guiding procedure guides the robot away from the goal to a new starting location at least 0.5 m away from the goal, mimicking the beginning of a new *trial* in rodent laboratory tasks.

All strategies and the gating network perceive the actions proposed by the guiding mechanism (there is no difference between an ordinary action and a forced action), so learning is performed in the usual way.

In the last experiments presented in this paper, once the robot has learned to reach the goal directly from its different starting position, the goal location is changed and the robot will adapt its behavioral policy to this new condition. Finally, once the robot has learned the new condition, the goal is moved back to its initial location in order to test the ability of the robot to restore previously acquired behavioral policy.

6. Results

In this section, we discuss the results we obtained on the Psikharpax platform. We will empirically show that our model can easily learn to associate a state with the best performing strategy in that state. The experiments were chosen to clearly show that the system works correctly (i.e. many of the results are predictable), instead of reproducing a complex protocol for which the evaluation of the quality of the model is inherently much harder to evaluate.

However, the experiments with multiple parallel strategies are in no way simple or unrealistic, given the limited and very noisy sensory input. Our results prove that with some small modifications, the Q-learning mechanism for the gating network from [33] works well on a real robot platform.

We first make a series of experiments to test the capacities of individual strategies (first planning, and then taxon) to learn the goal location and lead the robot toward it. In these experiments, the studied strategy is combined with the exploration (random) strategy in order to compare its performance with random movements of the robot, and to test the ability of the gating network to learn to stop selecting the exploration strategy when another strategy can lead the robot to the goal. In the next experiments, the taxon and planning strategies work in parallel and the gating network successfully learns which strategy is the most efficient in each subpart of the environment. In response to an environmental change (i.e. change of the goal location), the gating network manages to unlearn previous associations of strategies to subparts of the environment and to learn new ones, but this process is very slow. In a final experiment, we add the context-switching meta-controller to the system and show that it manages to adapt faster to environmental changes and to restore previously learned associations when a previously experienced context is again presented. Finally, we analyze in more detail a set of examples of cooperation between strategies produced by the system and allowing the robot to execute successful trajectories toward the goal.

6.1. Planning expert and exploration strategy

In this experiment, we connected two strategies to the gating network. The first one is the planning expert as introduced in

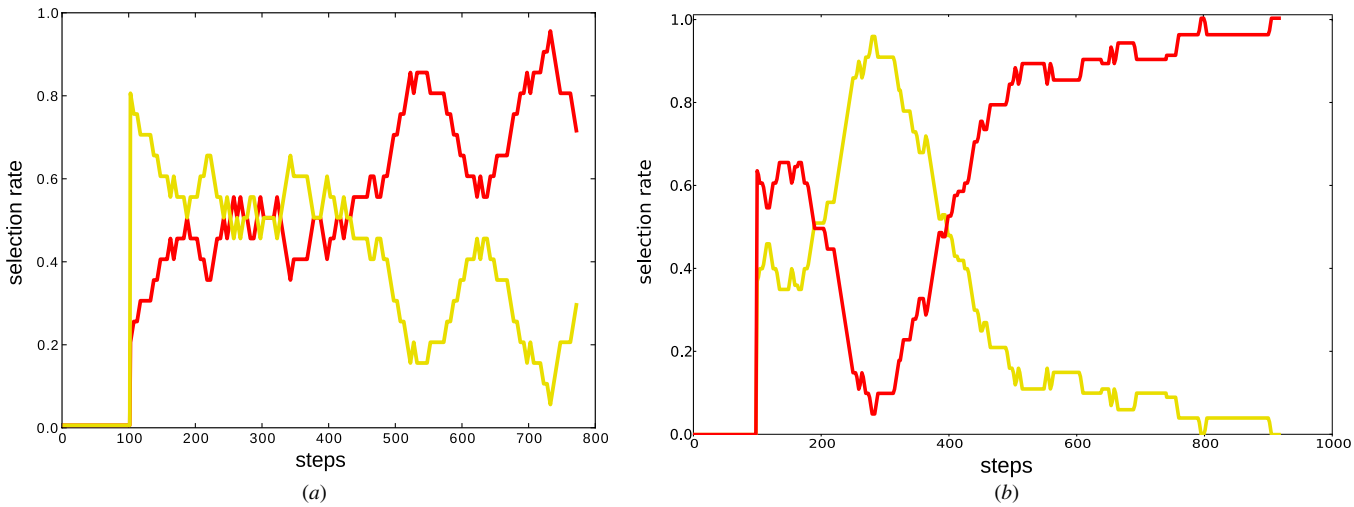


Figure 10. Selection rate of both strategies during learning. The dark line represents the planning expert, and the light line the exploration strategy. The horizontal axis indicates the current step (time). The transient phase ends for both experiments after about 300–400 actions. (*a*) Less precise topological map, resulting in an overall higher selection rate of the exploration strategy. (*b*) Very precise topological map; the planning strategy has very good performance.

the previous section. The other one is the random exploration strategy.

The goal of this experiment was to verify if the gating network could effectively learn to suppress a suboptimal strategy (the exploration strategy). Because the goal was relatively small and does not fall precisely on the center of activation of a place cell, the planning expert was only capable of efficiently guiding the robot to a zone around the goal. In other words, when the planning expert had reached a node in its map with higher diffusion values than any of its neighbors (given that the goal values are meaningful), the robot was not necessarily at a location where the reward is given, but only in the neighborhood. As a consequence, the robot's behavior remained random near the goal location—produced by a combination of the exploration expert and the planning expert which proposed random actions when it had attained the node in the diffusion vector with the highest rate.

Figure 10 shows the selection rate for the two experts as a function of the number of steps taken for two experiments (different environments and goal locations).

After an initial transient phase during which most of the Q-values were still small and meaningless, the system quickly converged to a regime in which the planning strategy was selected almost all the time. During this transient phase both strategies had similar performance because the diffusion values were not yet meaningful. We see that this happened for both experiments, but with a different (random) initial transient phase. However, the experiment from figure 10(*a*) which was slightly shorter than figure 10(*b*) and not yet fully converged had a higher selection probability for the exploration expert even after several hundred actions. After 300–400 actions (goal reached approx. five times), the gating network started to move toward its steady state. The Q-values would continue to increase but their relative values stayed stable.

To explain the observed difference in selection rates between experiments, we analyzed the locations in which each expert is the preferred strategy. This is illustrated by figures 11 and 12 where the size of the dots corresponds to the difference between the Q-values of each strategy at the mean position of each place cell: $\left|z_j^{\text{explr}} - z_j^{\text{planning}}\right|$. This corresponds to the selection probability in the gating network at each location. Larger differences indicate that the weights have differentiated more. The color/shape of each dot is determined by the strategy with the highest Q-values at that location (the winning strategy when following an optimal policy).

The planning strategy normally only needed one visit to the goal in a new environment to learn (i.e. to determine) good diffusion values. Figure 11 shows the evolution of the Q-values of the gating network for an experiment planning strategy versus exploration strategy. After 300 steps, the global structure no longer changed significantly. Close to the goal, exploration was often the preferred strategy, due to the coarseness of the planning strategy. Farther away, we see that the planning strategy was gaining terrain, because there the planning strategy performed well. The relative weights were still increasing (they diffused away from the goal due to the Q-learning algorithm), indicating that learning had not yet fully converged. In particular, the region at the lower-left corner still needed to be visited more to learn the Q-values in this region.

Note that in figure 11, there is a region around (1.5,0.8) where the exploration strategy remained competitive with the planning strategy for a long time. This is again a consequence of the topological map (shown in figure 8(*d*)). Such regions differed between experiments (different maps) and often indicated regions in which the topological map contained a detour. In such a case, it could indeed be a good strategy to transiently follow a random direction—as suggested by the *exploration* expert—to get onto another path where the planning strategy would be used again.

Figure 12 shows the result of another experiment with the same system (i.e. planning and exploration experts) but with different distal landmarks with less aliasing of the place cells around the goal and after a longer time of experimentation. In

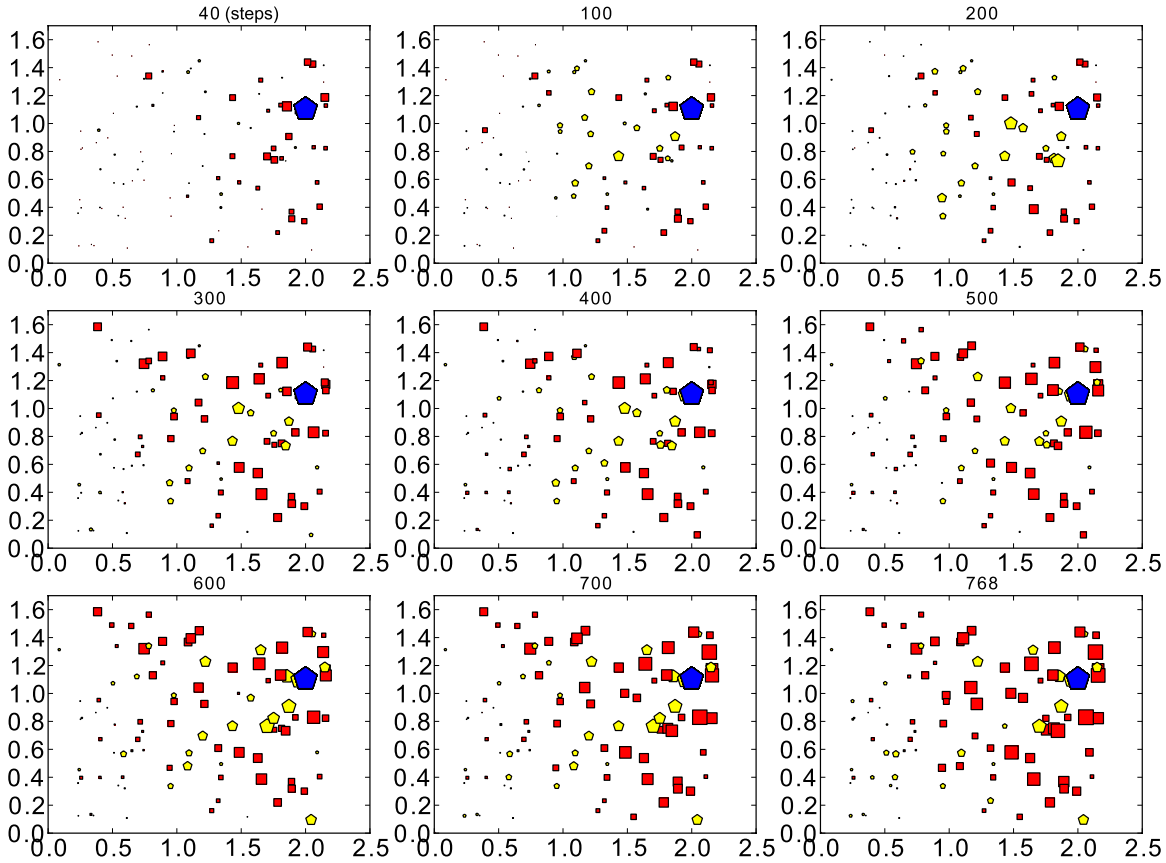


Figure 11. Evolution through time of the relative Q-values of the planning expert (dark squares) versus random strategy (light-colored pentagons) corresponding to figure 10(*a*). The relative Q-values $|z_j^{\text{explr}} - z_j^{\text{planning}}|$ are shown at the center of activation of each place cell. The dot takes the shape (square/pentagon) of the strategy with the highest Q-value (highest selection probability). The larger the dot, the more the weights have differentiated and the more likely the strategy with the highest Q-value is to be selected. The goal location is shown in blue. Competition between the strategies is still going on after 768 steps, but the structure becomes apparent. Near the goal, the random exploration strategy often performs equally well as the topological map. Axes in meters.

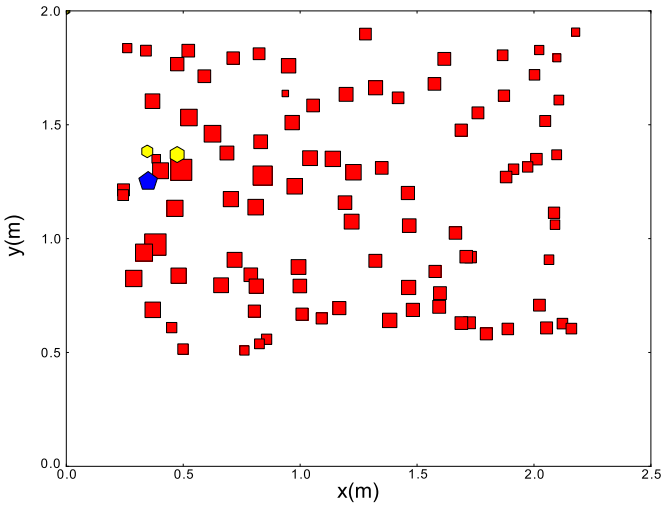


Figure 12. Planning expert versus random strategy corresponding to figure 10(*b*). Colors as in figure 11. The planning expert is almost always preferred, except very near the goal. Note the lower weight differences very near the goal.

this case, the gating network had stabilized, and the robot preferred the topological map strategy almost everywhere, except very close to the goal where the planning strategy was still unprecise and the robot's behavior was random. In the region near the goal, the weights of the gating network had less differentiated, indicating that the strategies are still competitive in this region.

6.2. Taxon expert and exploration

We now verify the taxon strategy by having it compete with the exploration expert. The setup is the same as in the previous experiment. However, because the taxon can only sense the goal within a certain range, the outcome we expected is different.

Once the taxon expert has learned to move toward the stimulus, its performance should be better than the exploration expert within a region around the goal. Farther away from the goal, the taxon does not receive sensory input and becomes equivalent to the exploration expert.

Figure 13 shows the evolution of the relative Q-values through time. It is clear that the taxon expert had significantly been reinforced around the goal and a large region around it. However, the taxon was still the dominant strategy relatively far away from the goal, which is against our expectations. We expected to see a more or less circular region around the goal in which the taxon was the dominant strategy, while the rest of the environment would be randomly assigned to one of both

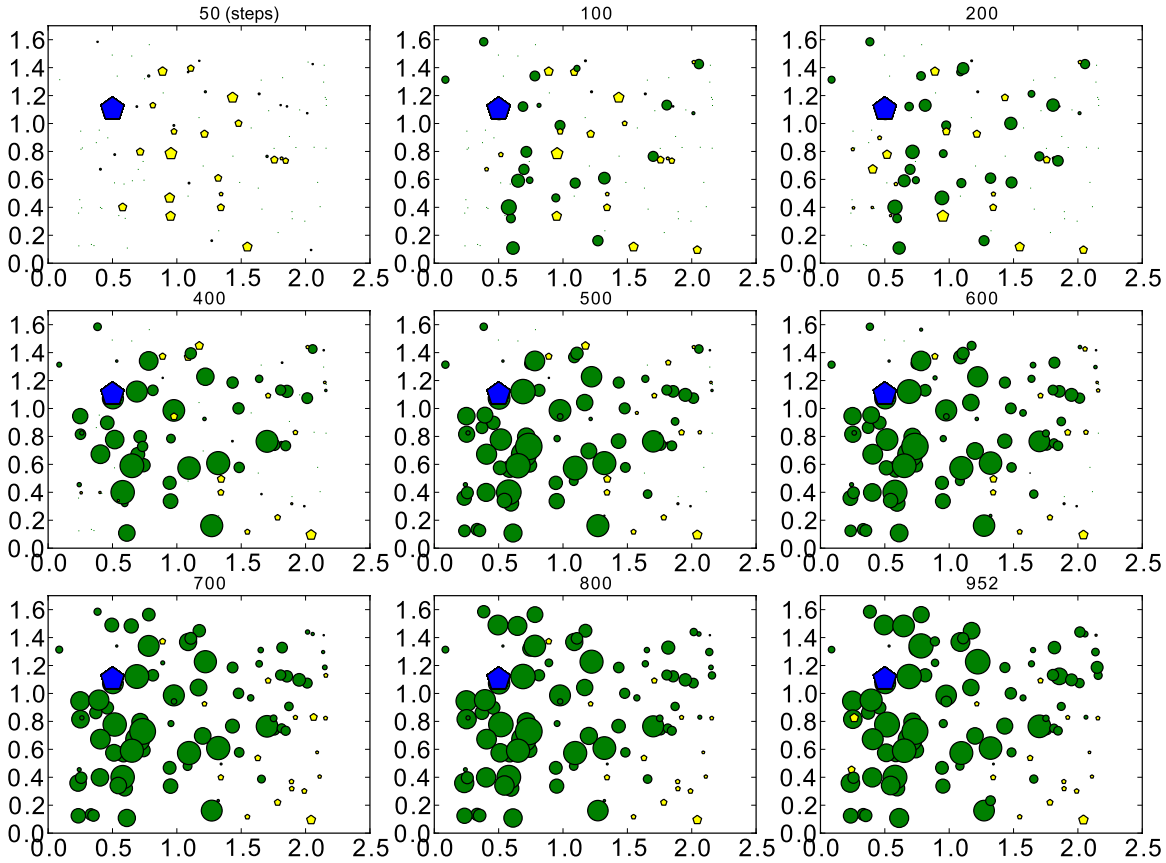


Figure 13. Taxon (dark circles) versus exploration expert (light pentagons). The goal is on the left (large dark pentagon). The gating network learned to select the taxon strategy near the goal, and preferred the random exploration strategy in some areas far away from the goal. Axes in meters.

strategies, resulting in very small weights on the plot at these locations (we plot the weight difference). While we indeed see this phenomenon at many locations (predominantly at the lower right side of the environment), there are regions far away from the goal with non-negligible weights for the taxon. This effect stems from the aliasing in the place cells. If some of the place cells far away from the goal were activated even slightly when the robot was close to the goal, then the optimal strategy around the goal (the taxon) was also slightly reinforced at these aliased locations. This was only of importance when none of the strategies performed well at the aliased locations farther from the goal, because otherwise this effect was dominated by reinforcements of these strategies. This indicates that the gating network gave a slight advantage to strategies having good performance around the goal when there was aliasing in the sensory input. One way to reduce this effect would be to only update the Q-values of the most active place cell in the gating network. This would increase the learning time significantly and the system would no longer use the similarity between nearby locations.

Despite such a limitation, the system appropriately learned to privilege the taxon expert near the goal and to select the exploration expert mostly on the right side of the maze, far away from the goal. Figure 14 shows the global selection rate of both strategies as a function of the number of actions. Although small changes continued to occur—due to the equal performance of both strategies farther away from the goal—the learning had mostly converged.

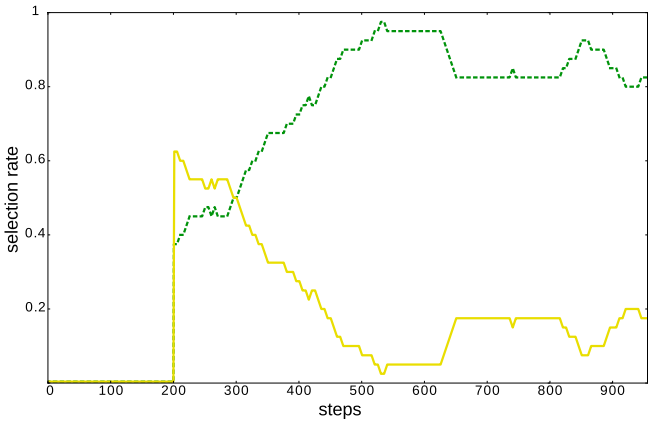


Figure 14. Selection rate of the taxon (dashed dark line) versus exploration strategy (light line) computed over the last 200 actions. The gating network clearly learned to privilege the taxon strategy.

Figure 15 shows the learned output direction of the taxon Φ^T for each input direction (see equation (4)). Here, we used a binary input vector r_j^{GP} to code the goal direction (the center of the Gaussian). The figure shows that the taxon learned to orientate toward the direction of the stimulus.

6.3. Planning expert and taxon (and exploration)

We then performed an experiment where we combined all experts and tested the ability of the gating network to select the right strategy in the right location. Based on the previous

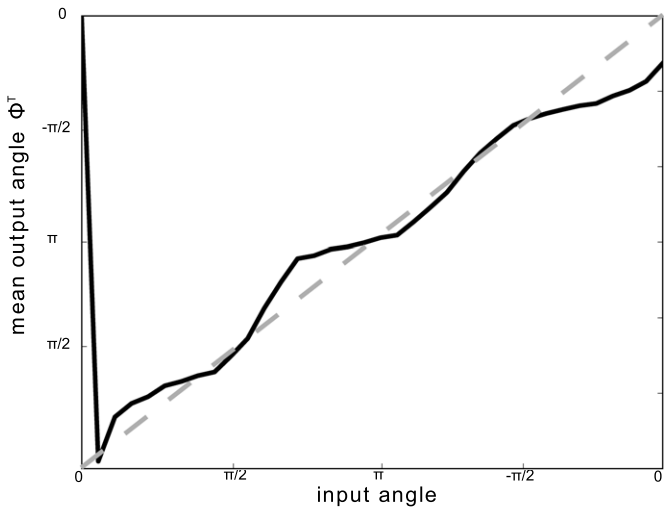


Figure 15. Output directions (weighted mean of angles) learned by the taxon in response to the input angle between the robot's head orientation and the goal. After learning, the taxon approximated a diagonal line and thus learned to orientate toward the direction of the stimulus.

results, we expected the animat to learn to prefer the taxon strategy when close to the goal, while using the planning expert farther away. The exploration strategy should normally be used at the beginning of the experiments, and then progressively be excluded in most locations, except in places with a high uncertainty or where there is indeed an advantage in following an arbitrary direction.

We conducted two sets of experiments. In the first set, the goal location was left unchanged and the robot started its exploration until converging to an appropriate coordination of strategies and adopting a satisfying behavioral policy to reach the goal. In the second set of the experiments, after a certain duration the goal was moved to a new location at the opposite side of the environment. This was done to evaluate the time required by the gating network to learn new associations of strategies with maze areas.

We performed two versions of this experiment. In the former, we connected only the taxon and the topological map strategy to the gating network. Furthermore, we pre-trained the taxon in a separate experiment (as in section 6.2) to study the complementarity of the two strategies once stabilized before testing the whole system. In the latter, we connected all three experts to the gating network. Each version used a different topological map, place cells and goal locations.

Finally, for experiments containing changes in the goal location, we tested the gating-network with and without the context switching mechanism enabled so as to illustrate its benefits on adaptation to environmental changes.

6.3.1. Part 1: fixed goal. We first discuss the results for the first version of the experiment: taxon and pre-trained taxon.

Figure 17 shows the result of the experiment after 30 trials of learning—the robot had reached the goal 30 times. The gating network had learned to choose the taxon strategy around the goal, as indicated by large weight differences in this area. Farther away from the goal, the weight differences were smaller and the planning strategy was preferred most of the time.

Figure 16(*a*) shows the selection rates of both strategies through time. We see that in short term, the selection rates varied (depending on the robot's trajectories) while the global selection rate had converged. Convergence was fast because each strategy performed well in a specific region and there was not much competition. Furthermore, the transitional phase was short because the taxon had been pre-trained (the taxon already had good performance around the goal).

Figure 18 illustrates the result of a session of the same experiment overlaid on an image of the environment to show the physical location corresponding to the different areas in figure 17.

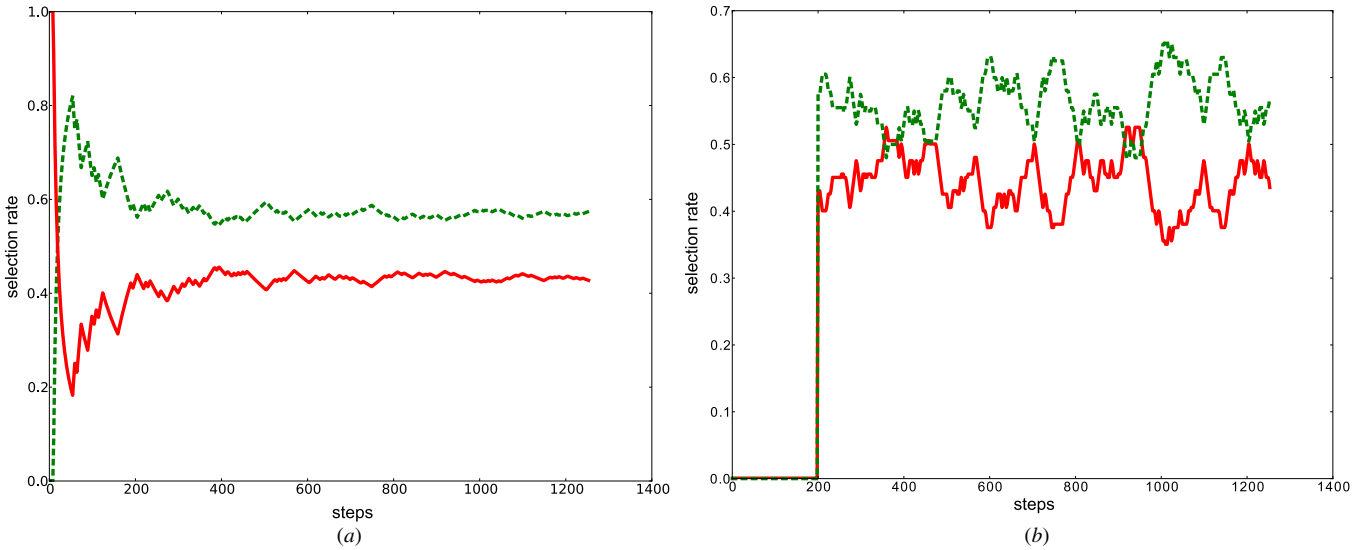


Figure 16. Evolution of the selection rate of both strategies (version 1) during learning (as a function of the number of actions). The full line represents the planning expert, the dashed line the (pre-trained) taxon strategy. The horizontal axis indicates the current step (time). The gating network globally converged to a stable repartition of selection of the two strategies, with the taxon being selected most often. (*a*) Selection rate computed over all previous actions. (*b*) Selection rate computed over the last 200 actions (moving average).

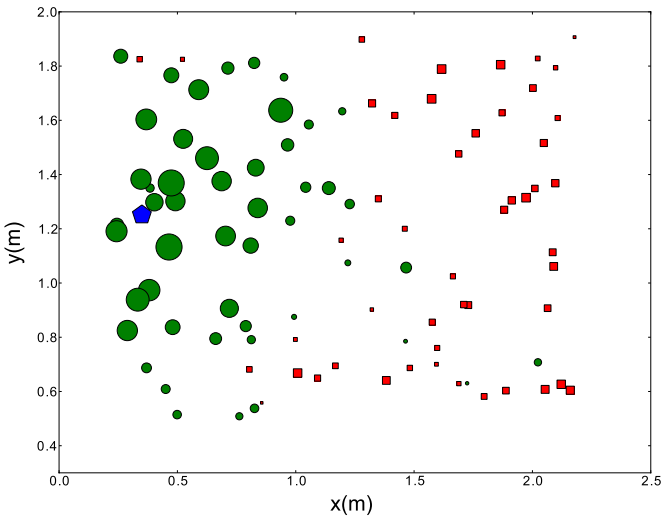


Figure 17. Planning expert (squares) versus pre-trained taxon (circles) (version 1). The goal location is shown as a dark pentagon on the left. The robot has learned to prefer the taxon strategy over the planning expert when close to the goal.

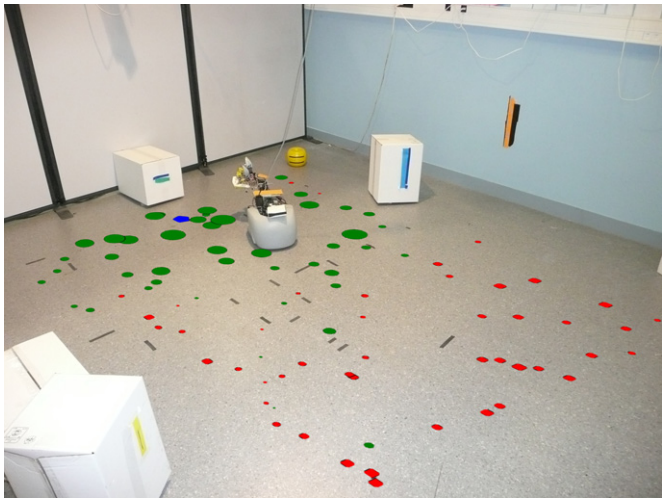


Figure 18. Result of the first version of the experiment illustrated in figure 17, this time overlaid on the environment.

For the second version of the experiment, we connected the taxon (this time not pre-trained), the planning and the exploration strategies to the gating network. Figure 19 shows the evolution of the relative Q-values through time and figure 20 contains the selection rates for the different strategies. This gives (as in the previous experiments) an indication of how likely a strategy was to be selected at a certain location and of how much the weights had differentiated. To show such differentiation between the three competing strategies, we plotted $\left|z_j^1 - \left(z_j^2 + z_j^3\right)\frac{1}{2}\right|$, where the indices $\{1, 2, 3\}$ correspond to the strategies ordered by the Q-value for the place cell j by descending order. In other words, we plotted the relative Q-value of the most likely strategy at each location compared to the average of the other strategies.

The figure shows that between 700 and 900 actions, the taxon took over the region around the goal, as expected. This was slightly longer than in the taxon-exploration experiment because the taxon needed to learn to move toward the stimulus and the gating network needed to learn the best out of three strategies. The topological map seemed to have performed very well in this experiment as even in a small region around the goal, this strategy was the winning strategy, although we expected the taxon strategy to win here. This effect is shown in detail in figure 21 which shows the relative Q-values at the end of the experiment for each expert separately (at the location where each expert was the most reinforced one). Clearly, the same structure as in figure 17 is found (without the exploration).

The exploration strategy's selection rate dropped to a significantly lower level than the other two strategies. The taxon and topological map strategies' rates approached their final state after 1400 actions. They continued to oscillate (depending on the robot's trajectory), similar to figure 16(*b*), but remained globally stable.

Now we still have to explain why the taxon performed badly in this small region mainly above the goal. Supposing that the gating network indeed worked correctly, this could only be due to the fact that the taxon indeed performed worse than the two competing strategies in this region. Now if we look at the learned directions of the taxon (figure 22), we indeed find that the taxon had not correctly learned the directions when the goal was located in the south ($-\pi/2$) of the robot.

Figure 23 shows the visited locations of the robot. It is clear that the robot had approached the goal (located at (2,0.9)) from the north very few times. Hence, the taxon did not have the possibility of learning to move toward the goal from above. The gating network learned that it was better to follow a random strategy (exploration or the planning when very near the goal) here, which was the best available choice.

Finally, there was a region around (0.5,0.5) where the topological map was not the preferred strategy in all cases. Figure 24 shows the topological map at the end of the experiment (step 1683). At the right of this region there was a large gap in the map (two disjoint trajectories). There were multiple nearby locations which pointed toward almost opposite directions (because the map tried to find a path with the least number of nodes). So if there was some aliasing in the place cells in this region, the topological map could oscillate here and thus have low performance, as we explained for the experiment with the topological map and exploration expert connected to the gating network. Figure 21 shows that the gating network learned to also select the other two strategies in this region to compensate the limitations of the planning strategy in this zone.

Globally, we got very similar results for both versions of the experiments. Once the robot had experienced the goal, the exploration expert's selection rate progressively decreased to a very low level. The main conclusion is that the gating network indeed worked as proposed and could easily learn to use the topological map when farther away from the goal and the taxon when closer. Interestingly, the system also solved the more complex situations in which aliasing of the map caused locally bad performance.

6.3.2. Part 2: changing the goal location. In the second set of experiments, after 1000 steps the goal was moved to

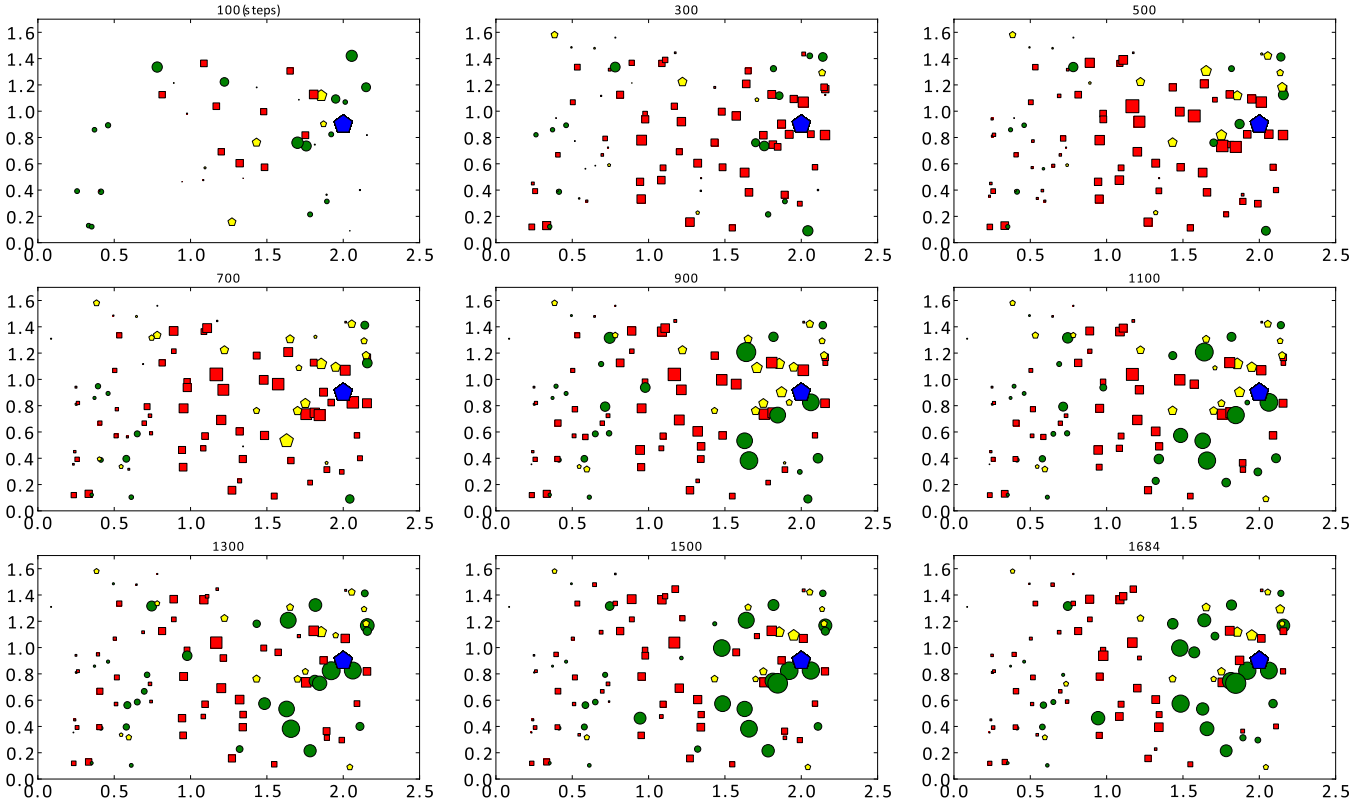


Figure 19. Evolution of the relative Q-values (see the text) for the second version of the experiment with taxon, topological map and exploration experts. The taxon strategy is preferred around the goal, while the planning strategy has higher Q-values farther from the goal (as in the simplified version). The exploration expert only wins in a small area (see the text for explanation). Taxon: dark circles; topological map: dark squares; exploration: light pentagons; and goal: dark pentagon on the right. Axes are in meters.

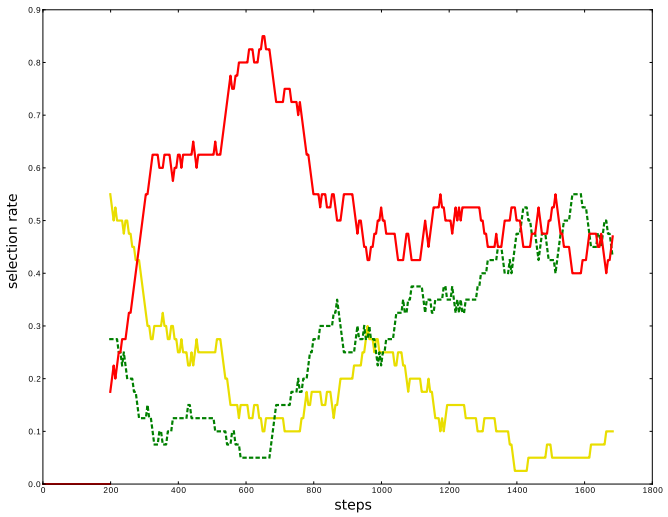


Figure 20. Selection rates computed over the past 200 actions (moving average) for all three experts. The taxon is shown as a dashed dark line, the topological map as a dark full line and the exploration expert as a light line. After an initial learning period, the taxon increases in importance (approx. 700 actions) and finally settles in a similar regime as in figure 16(*b*). The exploration expert has a very low selection rate as expected.

the opposite side of the environment to test the ability of the system to adapt to a new situation. We first present results with the context-switching mechanism disabled.

The experiment was relatively difficult as the only visible change to the gating network was the reward. The result from a session (first version) is shown in figure 25. The gating network appropriately learned to stop selecting the taxon expert on the left side of the maze and to rather select it on the right side of the maze (near the new goal location). Thus, the robot adapted to the new situation. However, this process was very slow and took approximately 8000 steps (approx. 180 trials). Even then, the taxon remained the preferred strategy in the area around the old goal. This is because the robot needed to unlearn the previous associations between strategies and maze areas before learning the new ones.

Rats typically note drastic changes in the environment and can adapt their behavior fast when they are not overtrained—in which case they build unflexible habits [17]. In our own previous strategy shifting experiments with real rats, in response to task changes, animals abandoned the previously performed strategy after an average of ten error trials and learned to select the new appropriate strategy in about 100 trials [8, 66]. Thus, the gating network alone is not sufficient to produce behavioral performance and adaptability comparable to real rats.

6.4. Context switching

To overcome this limitation (inherent to the Q-learning algorithm), we implemented a simple context-switching mechanism which associates with each different goal location a different context and thus a different instance of the gating network (see section 4.3). The idea was to anchor the detection

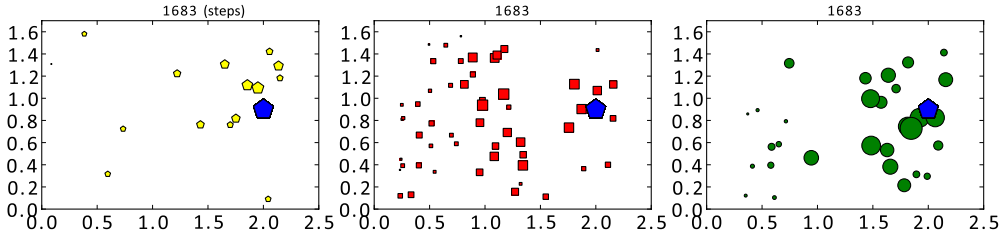


Figure 21. Left: exploration; center: topological map; right: taxon. Relative Q-values for each expert at the end of the experiment at locations where each strategy has the highest Q-value (highest selection probability). Axes are in meters.

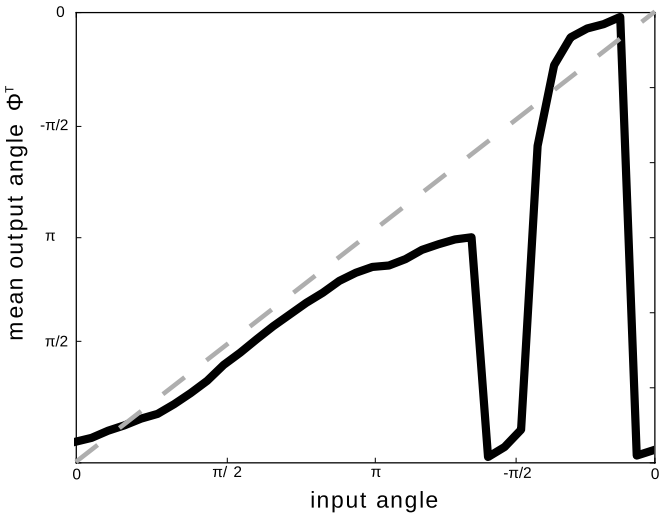


Figure 22. Learned orientations of the taxon. Each time the goal was located at the south of the robot ($-\pi/2$), the taxon did not learn to move toward the stimulus.

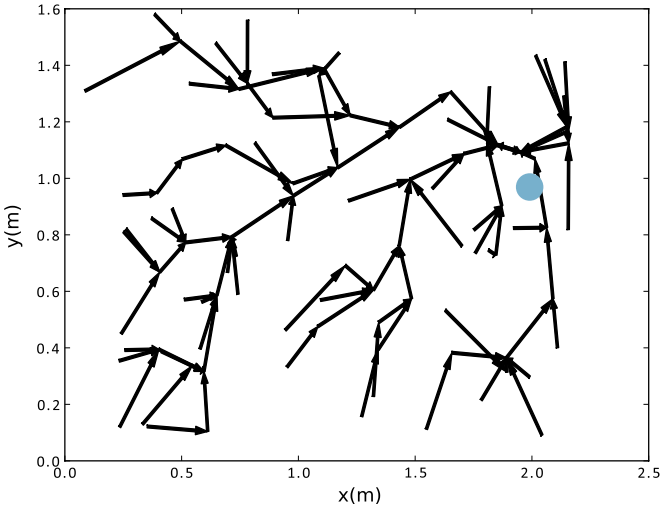


Figure 24. Topological map after 1683 steps. The goal is shown as a circle on the right.

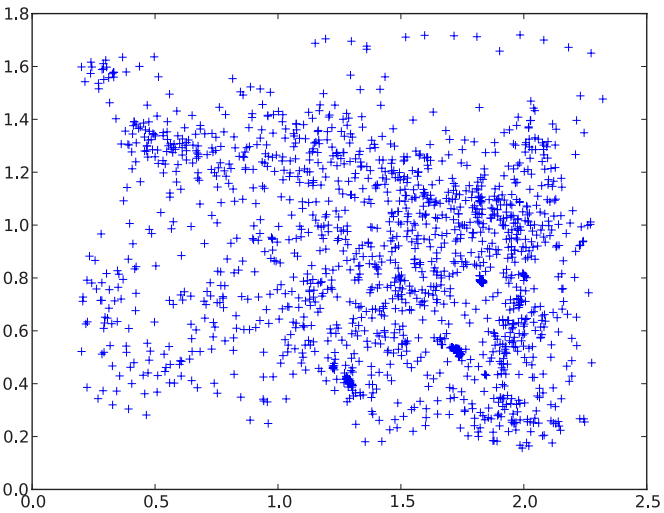


Figure 23. Visited locations of the robot. The goal was located at (2.0,0.9). The region right above the goal (the goal extends up to (2.0,1.0)) had been visited much less by the robot, which explains why the taxon had not learned the correct response when the goal was situated at the south of the robot. Axes are in meters.

of new contexts in response to a change in goal location on the diffusion values in the topological map of the planning expert. While the place cell activation (i.e. the input of the gating network) does not change when the goal is moved, the goal values g_j in the topological map do. However, the goal values can fluctuate heavily and it is better to use the diffusion values d_j. The topological structure was thus used to compute the diffusion values, which gave them a much smoother activation.

As in the first set of experiments, we first discuss the results of the first version (no exploration expert).

The previous experiment was repeated with this mechanism (figure 26). In total, four contexts were recruited (two transitional). The goal was moved twice to verify that the robot had stored the initial context and that it could appropriately restore it. After 900 steps, the goal was moved from the initial location on the left to the new location on the right. Once the robot found the new goal location, the topological map automatically produced different diffusion values and the system could create a new context (illustrated by the abrupt vanishing of most circles and squares at step #1100 in figure 26). After 200 more steps, the gating network had learned to select the taxon expert near the new goal location and the planning expert in most of the rest of the maze. At step #1450, the goal was moved back to the initial location. Because the contexts had been stored, the robot recalled the previously learned weights when the goal moved back to its initial location. This resulted in instantaneously restoring previously learned weights of associations of the gating network (illustrated at step #1770 in figure 26).

In total, the robot had constructed four contexts. The first one corresponded to the initial phase in which the robot had yet to find the goal for the first time. The second one was the context used to learn when the goal was on the left and when the

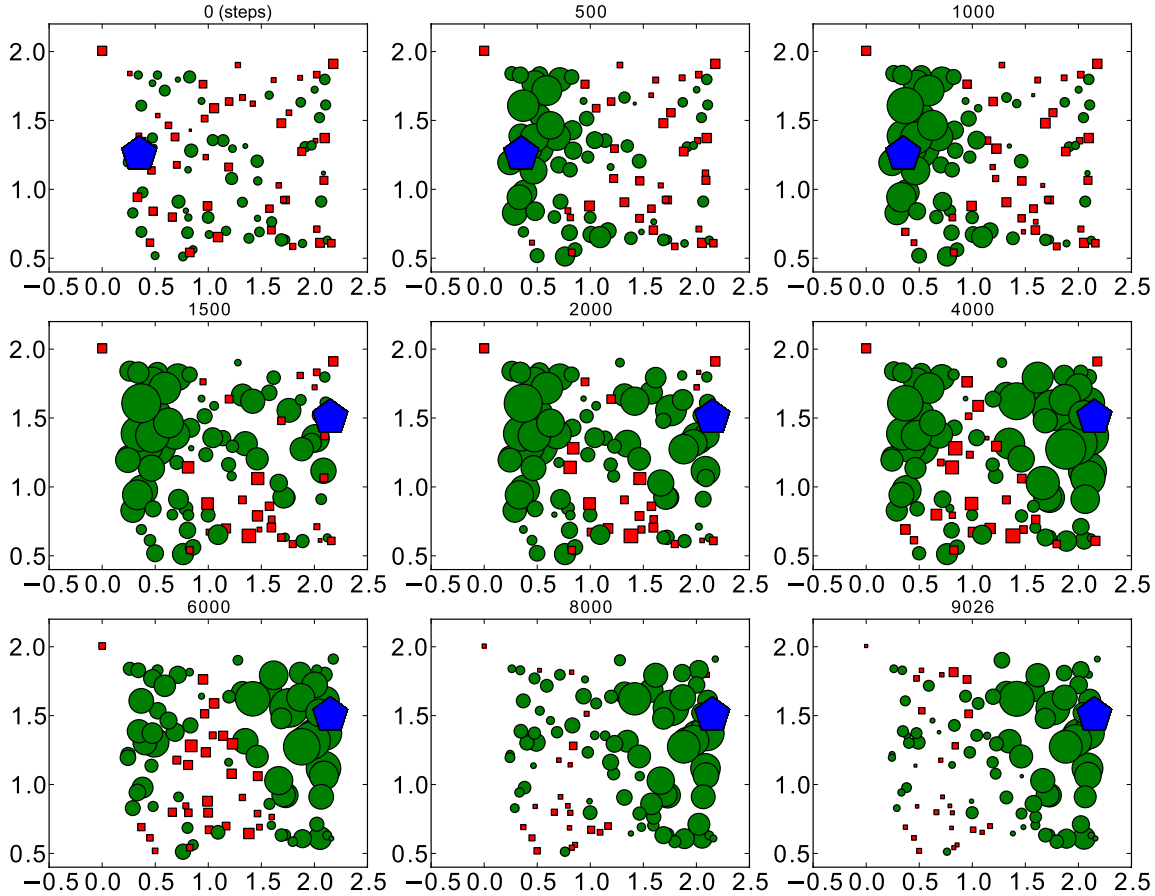


Figure 25. Planning expert (squares) versus taxon (circles) (first version of the experiment). The goal location is represented by a dark pentagon. The location of the goal is changed after 1000 steps. The gating network appropriately learned to stop selecting the taxon expert on the left side of the maze and to rather select it on the right side of the maze (near the new goal location). However, this process was very slow and took approximately 8000 steps, corresponding to around 180 trials. Axes in meters.

goal was moved back to the left (at the end of the experiment). The third one was a transitional context used when the goal had only recently changed sides. This is probably due to the change in diffusion values of the topological map once the robot did not get reward at the initial goal location, although the robot had not experienced the new goal location yet. The last one was used when the goal was found on the right side of the maze.

With this simple extension, learning was much faster and the robot could recall previous contexts which made the navigation system much more useful in practice. Another small but useful extension would be to associate a ζ^i with each context. This way the robot could approach the winner-take-all strategy for contexts that had been learned for a longer period.

We repeated this experiment (moving goal with context switching) with the second version (all three experts). As the gating network learns more slowly when three strategies are present, we moved the goal after 1683 steps (we continued the experiment from part 1).

Figure 27 shows the selection rate of the three strategies. Because the context-switching mechanism recruits a new set of Q-values when the change is detected, learning in the gating network can restart from scratch, which is faster than when the same context is used.

In figure 28, the average Q-values for each expert are shown (averaged over all locations). When the context switch occurs, the weights become zero (new context). Finally, figure 29 shows the selected context at each step. In total, again four contexts were created of which two were transitional. The context selection mechanism is stable (this depends on the threshold).

6.5. Cooperation: typical trajectories

To prove that the robot could learn to make strategies cooperate, we compared a number of trajectories of the robot equipped with only the planning strategy or with both the planning strategy and the taxon strategy. The goal was not moved during the experiment.

To illustrate abrupt switching between strategies when both strategies were used, the gating network was used with ζ initially set to 1 and to ∞ after learning (the optimal strategy always wins).

Figures 30(*a*) and (*b*) show two typical paths taken by the robot to reach the goal. Figure 30(*a*) is a trajectory obtained when only the planning strategy was used. Due to the limited precision of the topological map (place cells), the robot took pseudorandom actions near the goal. It could not plan a path leading closer to the goal. Figure 30(*b*) shows the cooperation

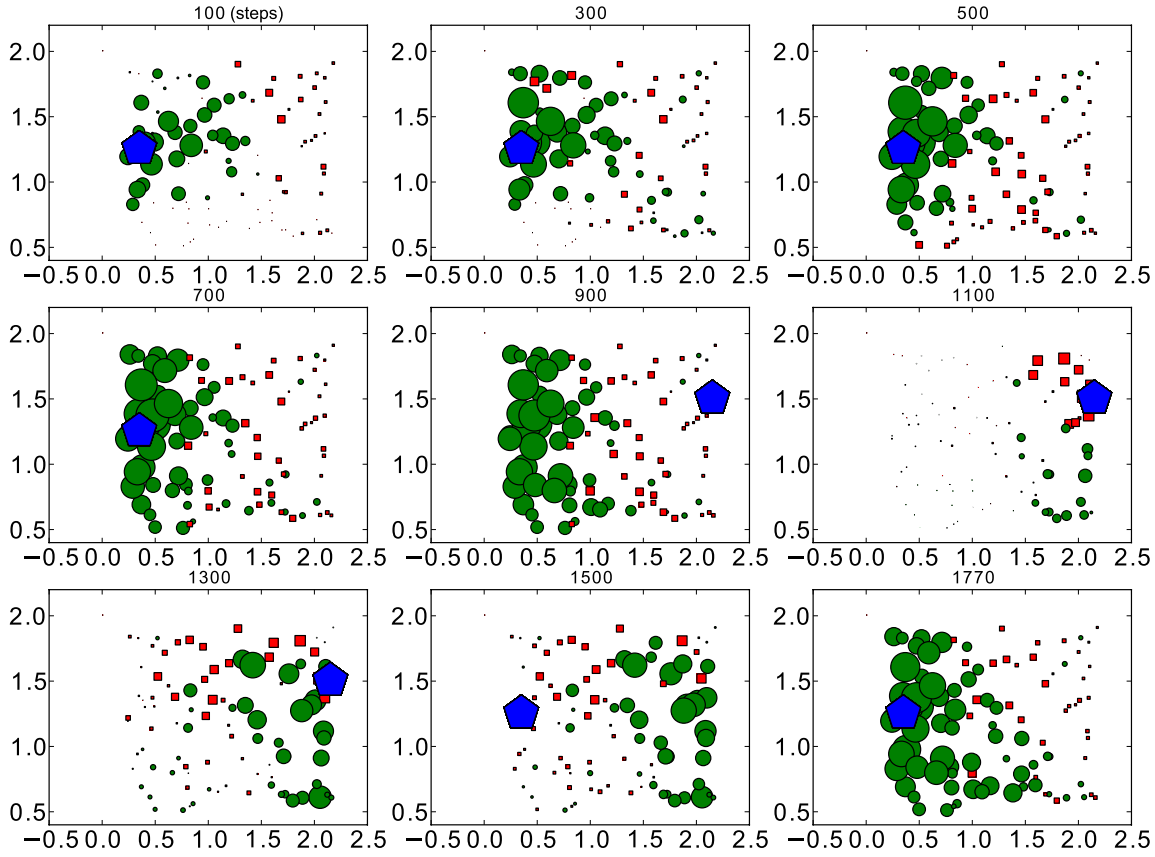


Figure 26. Planning expert (squares) versus taxon (circles) with the context-switching mechanism. The goal location is represented by a dark pentagon. Learning is now faster as the robot recalls previously learned contexts and new ones. The number of steps is shown at the top. With the context-switching meta-controller, the robot could quickly adapt its behavioral policy to a change in goal location, and could quickly restore its initial policy when the goal was moved back to its first location. Axes in meters.

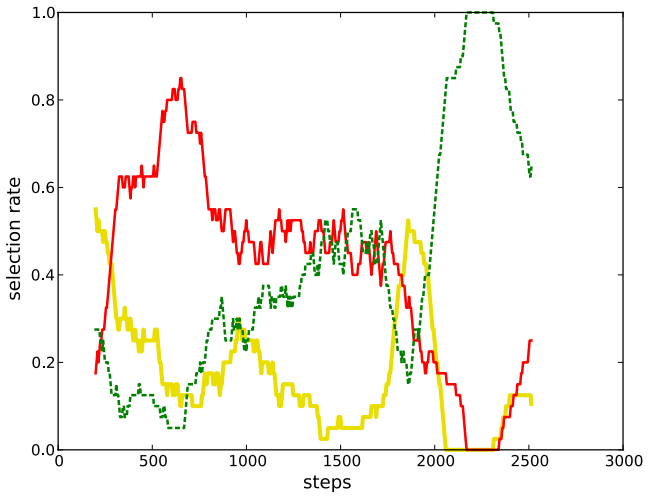


Figure 27. Selection rates computed over the past 200 actions for all three experts. The taxon is shown as a dashed dark line, the topological map as a dark full line and the exploration expert as a light line. The goal was moved after 1683 steps and shortly afterward we see a new transient phase in the selection rates because a new context was created.

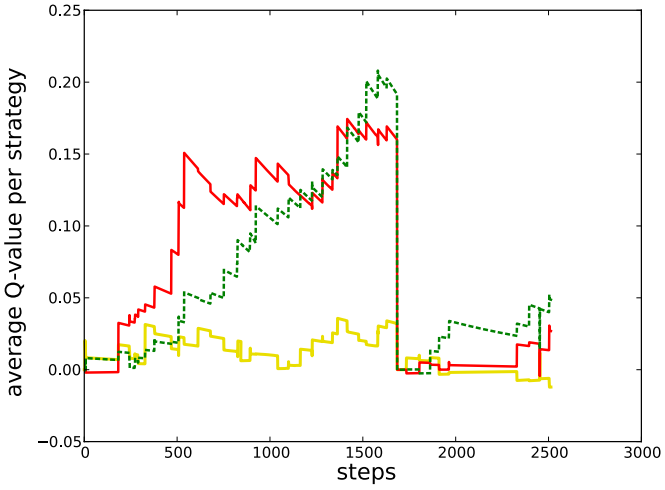


Figure 28. Average Q-values (over all locations) of the current context for each of the strategies as a function of the number of actions. Shortly after the goal is moved, the weights become zero, because a new context is activated. Learning restarts in this new context. Colors as in figure 27.

between the taxon and the planning strategy. The robot had learned that when it was near the goal, it should rely upon the taxon strategy as it could guide the robot when the goal was close enough[8]. When only the taxon strategy was used (not shown) and the robot was placed far away from the

[8] A video illustrating such coordination of navigation strategies in the robot in a simple example is shown here: http://chronos.isir.upmc.fr/~khamassi/projects/Psikharpax/VideoCaluwaerts2010.mov

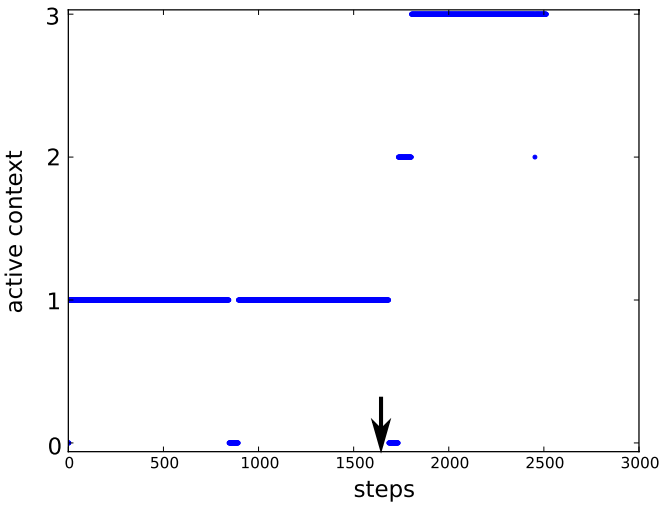


Figure 29. Active context through time. When the goal changes location (indicated by the arrow) a transitional period starts and until the diffusion values have adapted to the new goal location (context 3).

goal, the robot did not approach the goal and instead moved randomly.

7. Discussion

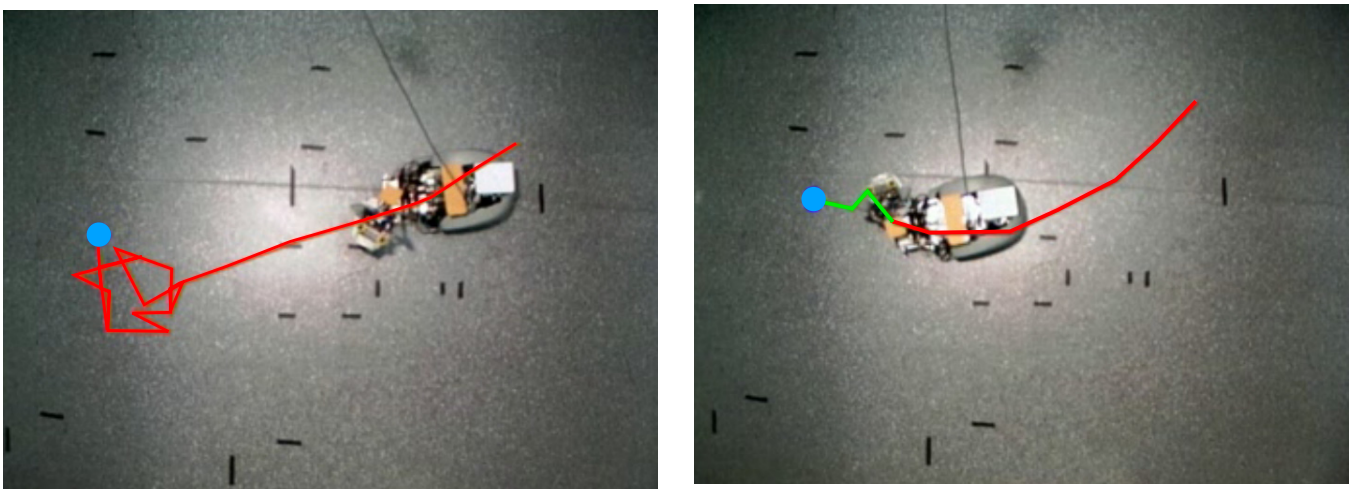
We presented the implementation of a novel strategy selection meta-controller allowing an autonomous robot to navigate an initially unknown environment. The model selected among two parallelly learned navigation strategies: a response strategy learning directions of movements in response to perceived cues; a place strategy building a map of place cells and planning trajectories between different maze areas. The strategy selection meta-controller was added to a previously published model of multiple navigation strategies [33] which was tested in simulation to replicate a series of rat behavioral experimental results [34, 18].

It was shown that the animat can learn to ignore useless strategies very quickly (e.g. an exploration expert after the goal was found). Furthermore, the robot learned to associate states with optimal strategies even in more complex cases, when for example a local taxon strategy was combined with a global but coarse path planning strategy. By introducing a simple context-switching mechanism, the robot could adapt quickly to changes in the environment.

The results are encouraging in two ways. First, the simulation model was adapted and validated on a real robotic platform, which was the main goal of this work. For the model to work on a robotic platform, the sensor input and model parameters had to be adapted. Because of these differences, the results presented here cannot be compared directly with the simulation model in the quantitative sense. However, the general principle of *common currency* was shown to be flexible and allows us to investigate the behavior of our robotic rat with different combinations of strategies in a realistic setting.

Second, the robotic experiments revealed that a simple context-switching mechanism can drastically increase the performance. Such a mechanism was absent from the simulation model as the robot could learn different situations with a single gating network because more sensory inputs about perfectly distinguishable landmark cues were available. Hence, our context switching shows that good performance can be obtained with less sensory input (only place cells feed into the gating network) and it has the added benefit of long-term storage of situations.

In terms of neural substrates, such meta-controller may constitute a model of rat prefrontal functions during strategy-shifting. Indeed, although less differentiated than primates' PFC, the rat PFC is known to have strong functional homologies with the lateral PFC in primates [67]. It is important to achieve high-level cognitive processes, usually referred to as executive functions, that is 'complex cognitive processes required to perform flexible and voluntary goal-directed behaviors based on stored information in accordance with the context' [68]. Responses of rat PFC neurons to working-memory components [69], spatial goals [70] and action-outcome contingencies [71, 72] initially suggested to several authors that the rat PFC could be the neural substrate for a particular behavioral strategy, the planning system or more generally for model-based learning processes—that is, decision-making based on the learning of transitions within the environment by means of action-outcome contingencies [50, 48, 49]; see [8] for a review. However, lesions of the rat PFC only impair the acquisition of goal-directed behaviors—that is, model-based strategies such as planning [73]—but not their expression [74]. Besides, lesions of the rat PFC impair working-memory processes only when combined with other factors such as the difficulty of the task, attentional mechanisms or the requirement for flexible behaviors [68]. This suggests that the neural substrate for the planning navigation strategy is located elsewhere in the brain, and that the rat PFC might be involved in a higher level of decision making and adaptation [8]. Consistent with our interpretation, on the one hand, neural correlates of forward planning have been found in the hippocampal system [75] and projections from the hippocampal system to the ventral and dorsomedial striatum appear important for model-based learning such as the quick adaptation to changes in the association between places and rewards [23, 76, 8, 77, 78]. On the other hand, PFC was found to be crucial for switching between behavioral strategies in response to task-rule changes [35–37, 79, 80]. Neural responses of the rat PFC show abrupt changes when the animal switches its navigation strategy [81, 82, 66], and only neurons responding for the correct strategy—i.e. the rewarded one—are reactivated during sleep in interaction with the hippocampus, therefore contributing to the consolidation of the association of the right strategy with the right context [66]. Our present meta-controller—combination of a gating network that learns to associate strategies with subparts of the task and a context-switching detector—constitutes a proposition of how such strategy-shifting functions may be modeled. We found that during robotic tests in the real world as opposed to previous simulations of the model, such a system was required on top of the planning and taxon strategies to produce fast adaptation to task changes. Such meta-controller may correspond to a minimal form of cognitive control models, which are used



(*a*) (*b*)

Figure 30. Sample trajectories of the robot: (*a*) only planning strategy enabled, (*b*) planning and taxon strategies enabled. (*a*) Planning strategy only: the robot approaches the goal quickly, but cannot reach it efficiently due to the limited precision of its topological map. (*b*) Planning and taxon strategies: the robot approaches the goal with the planning strategy as before and switches to the taxon strategy at the end to precisely reach the goal.

to model the role of primates' PFC in memorizing which task sets are relevant in which contexts [65]. In future work, we plan to test the robot for artificial lesions of this meta-controller, as compared to lesions of the planning system only, to compare with rodents behavioral data previously obtained during various maze tasks commonly used in the neuroscience community.

Finally, this work also has the potential of contributing to mobile robotics. Indeed, the bio-inspired ability to rapidly switch between several behavioral strategies and to memorize which strategy is the most efficient and appropriate in each subzone of the environment could help improve current control architectures for robots. Multi-layered control architectures with different levels of decisions have become more and more popular in robotics and are now widely used [83–85]. Such architectures open issues such as managing the interactions between submodules, coordinating multiple competing learning processes and providing alternative solutions to motion planning in situations where such a strategy is limited [86]. Indeed the planning strategy can be approximative when coping with uncertainties, e.g. when there is perceptual aliasing as we have seen here, and can also require high computational costs and a long time to propagate possible trajectories through mental maps [84]. In contrast, in situations where animals have developed habits under the form of cue-guided taxon or response strategies to solve a particular task, they can perform quick and accurate decision making. Moreover, in the case of a sudden change in the environment, they can adaptively abandon habits in favor of planning new routes toward their current goal [87]. Taking inspiration from computational models of how the mammalian brain learns to select appropriate strategies and to switch between strategies as a function of a speed-accuracy trade-off may constitute the basis of great future advances in robotics [73, 88].

Acknowledgments

This research was supported by a European Commission grant to the FP6 project 'Integrating Cognition, Emotion and Autonomy' (ICEA, IST-027819, www.iceaproject.eu) as part of the European *Cognitive Systems* initiative. KC was funded by a PhD fellowship of the Research Foundation—Flanders (FWO). MS was funded by a PhD fellowship from the EC FP7 DEXMART project.

Appendix A. Visual processing and place cells

A.1. Visual system

An overview of the complete visual system is given in figure A1. At the top, the BIPS hardware—bio-inspired perception system hardware processors [4]—tracks objects within the field of view of both cameras. Figure A2 gives an overview of the BIPS hardware. The visual input (15 fps) is passed through a set of elementary filters, similar to those in the primary visual cortex and the prestriate cortex. Next, the so-called tracking units compete through inhibitory connections to track objects based on feature coherency, similar to the extrastriate cortex. A simple matching procedure is used for stereoscopic vision. Objects tracked by the dominant eye (camera) are matched with objects seen by the other camera based on their location.

The elementary features such as shape, orientation and size are coded using neurons with Gaussian activation functions for each direction within the field of view in layer L2 of A1. As explained in section 2.2, the combination of this information with the odometric system results in panoramic information of visual cues around the robot. It is important to note that the robot does not identify landmarks in this system, and it only uses a constellation of detected objects.

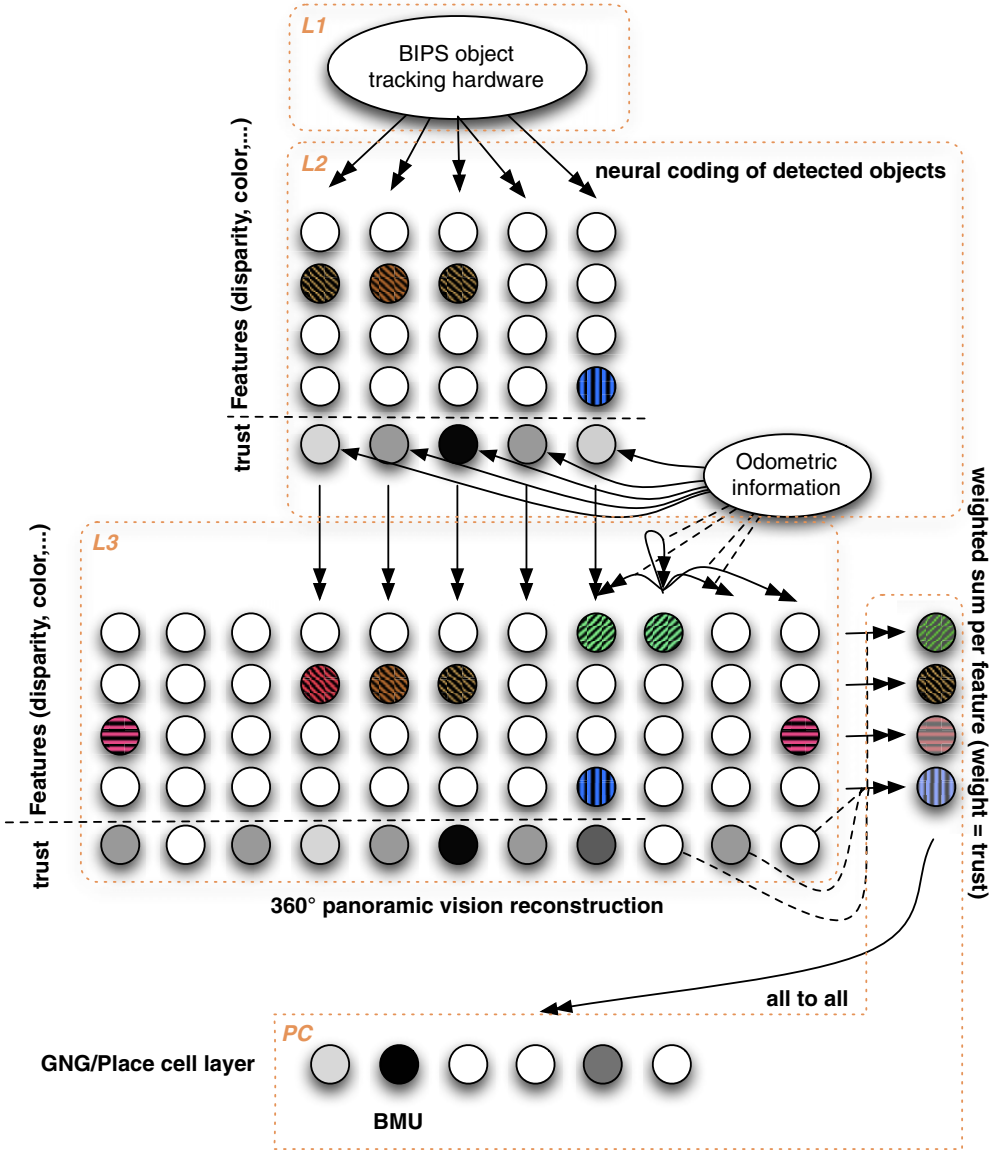


Figure A1. Overview of the visual system. L1: first layer of the visual system, consisting of the bio-inspired perception system hardware to detect and track objects. L2: in this layer the features of the objects from L1 are coded on a set of neurons. It only contains information on the objects within the field of view. L3: this layer contains a memory of all features around the robot (360 degrees). The L2 layer projects onto this layer to update the part of the memory currently within the field of view. The odometric system is used to modulate the projections from L2 to L3 to prevent updating L3 when L2 is expected to be unreliable. L3 is egocentric and hence there are lateral connections between neurons of different directions which are modulated by the odometric system (rotation of the head). PC: finally, the features from L3 are averaged (weighted) over all directions and project onto the place cells. BIPS: bio-inspired perception system [4]. GNG: Growing Neural Gas [44]. BMU: best matching unit.

A.2. Visual data clustering for place cell building

Different techniques with biologically inspired equivalents are available: Hebbian-like learning rules, principal component analysis [42], independent component analysis [89], SOMs [43], etc.

We initially tried linear approaches such as principal component analysis [42] to check if the inputs were linearly separable, but the number of regions (place fields) that could be recognized was too low (the precision of the place cells would be too low to be usable).

A popular nonlinear alternative for clustering are SOMs [43]. In a SOM, a fixed number of neurons is used with a predefined topology. Normally a two-dimensional topology is used, so the SOM performs an N-to-2-dimensional mapping.

The goal of the SOM is to move the neurons in the high-dimensional input space to approach the topology of the input. For each input, the Euclidean distance between the input and each neuron of the SOM is computed. The closest neuron is called the BMU (best matching unit). The SOM is now updated by moving the BMU closer to the input (weighted sum) as well as its neighbors. One can use a fixed neighborhood or a decay factor for this.

SOMs are conceptually simple and they are a very powerful tool to discover clusters in a dataset. A first disadvantage of SOMs is that the number of neurons is fixed.

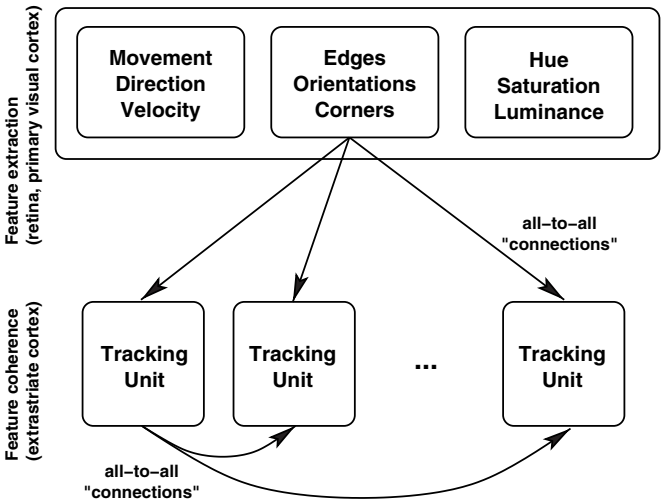


Figure A2. Details of the BIPS system [4]. First elementary features are extracted and then objects are tracked by the tracking units.

This problem can be overcome by trying out different sizes and visualizing the result. A more important problem is that the topology is fixed, i.e. the neighbors of a neuron stay the same. This problem is difficult to overcome when the topology of the input is highly irregular.

A last problem is that one typically chooses a large number of neurons to create clearly distinct zones in the output. While this is not a disadvantage for visualization, one would prefer a small number of output neurons (i.e. place cells) in which each neuron codes for a distinct zone in the input space.

To overcome these problems we moved to a different and closely related type of artificial neural network, the so-called GNG [44]. GNGs are created incrementally by inserting a new neuron after a number of input samples by splitting the neuron with the largest accumulated error (sum of distances) into two new neurons.

The topology itself is also learned by keeping the neighborhood of the neurons up to date. GNGs are generally better at approximating input topologies with high-error zones than SOMs, because the topology is not kept fixed and neurons are placed where they are most useful. A SOM can be seen as a predefined graph of which only the value of each node is updated (their position). A GNG also learns the graph by inserting nodes and edges. Both SOMs and GNGs are vaguely similar to the classical k-means algorithm [90], but the update rules are local.

We used the modular toolkit for data processing implementation [91] of the GNG algorithm with the default parameters. The distance measure used was the Euclidean distance, as decorrelating the input variables with the Mahalanobis distance [92] did not improve the quality of the results.

We now compute the input to the GNG layer as follows, where d again refers to the orientation of the neuron, j to a feature and $j = N_{\text{feat}}$ to a trust neuron:

$$u_j^{PC}(t) = \frac{\sum_d n^{L3}_{d,N_{\text{feat}}}(t) n^{L3}_{d,j}(t)}{\sum_d n^{L3}_{d,N_{\text{feat}}}(t) n^{L3}_{d,0}(t) \ldots \sum_d n^{L3}_{d,N_{\text{feat}}}(t) n^{L3}_{d,N_{\text{feat}}-1}(t)}. \tag{A.1}$$

This equation is valid for all features except for the confidence neurons, which do not project onto the GNG layer (they only modulate the other inputs).

Because the GNG algorithm inserts a new neuron after a fixed number of samples, the number of neurons would grow out of bounds after a while. To solve this problem, we fix the maximum number of neurons. This way, the animat creates new neurons at the beginning of the exploration and is forced to reuse the existing neurons when the maximum number of neurons is reached.

For a sample to be considered by the GNG layer for learning, the mean trust level $\left(\sum_j n^j_{\text{conf}}(t)/N_{SC}\right)$ must be greater than 0.75. This limit was found empirically as the quality of the place cells stopped improving above this value. A limit of 0.75 means that the robot needs to be able to see about 270 degrees of its environment from time to time. This is mostly due to the fact that we are using an open environment with resembling cues and a very noisy and unstable input. In a labyrinth with distinct cues in the corridors, we estimate that the robot can navigate with much less information.

Another solution to enable navigation with a lower mean trust level is to use a GNG layer with much more neurons and to perform a longer exploration. This way the robot will learn orientation-dependent place cells. The problem is that the planning strategy needs to learn much more connections, because multiple representations will exist for the same place.

A.3. Place cell activation

The activation of the place cell is computed as follows:

$$u_k^{PC} = \frac{\sum_{o=1}^{O} n^{L3}_{\text{conf},o} n^{L3}_{\text{feat},o,k}}{\left\| {n^{L3}_{\text{conf}}}^T n^{L3}_{\text{feat}} \right\|} \tag{A.2}$$

$$\Delta_j = \| v_j - u^{PC} \| \tag{A.3}$$

$$r_j^{PC} = \frac{\Delta_j^{\nu}}{\max_i \Delta_i^{\nu} \| \Delta \|}. \tag{A.4}$$

Here v_j are the weights of the jth node in the GNG (the jth place cell). The index o stands for orientation (each orientation has a set of feature neurons and a trust neuron). $n^{L3}_{\text{conf},o}$ is the oth element of the vector (one element per direction) of the vector of trust neurons n^{L3}_{conf}. Similarly, $n^{L3}_{\text{feat},o,k}$ is the neuron of the kth feature for direction o. Hence, n^{L3}_{feat} is an O by F matrix with O the number of orientations and F the number of features. We used $O = 960$ to prevent aliasing in L3 when the robot turns its head. As we sum over all orientations, the overhead of a large O only influences L3. ν defines the smoothness of the place cell activation and is a tuning parameter.

Throughout the main text, we use $\bar{n^{L3}_{\text{conf}}}$ to refer to the mean trust. This is defined naïvely as

$$\bar{n^{L3}_{\text{conf}}} = \frac{\sum_{o=1}^{O} n^{L3}_{\text{conf},o}}{O}. \tag{A.5}$$

Appendix B. List of parameters

Name	Value	Explanation	Remarks
N_{PC}	100	No of place cells	Arbitrary, but limits the precision of the gating network/topological map.
N_{PFC}	98	No of nodes in topological map	Arbitrary, but here fixed to N_{PC}, minus the place cells that are never the BMU.
N_{ANG}	36	No of direction neurons	Per connection in the topological map.
N_{DIST}	35	No of distance neurons	Per connection in the topological map.
β	0.75	Min. trust for map/path update	
ι	0.7	Goal diffusion factor	
ζ	1 or ∞	Selection probability exponent	Lower values speed up learning, higher values make the robot follow an optimal policy.
λ	0.76	GN eligibility traces decay factor	
ξ	0.05	GN learning rate	
γ	0.8	Future reward decay factor	
τ_{forget}	0.02	Reward location decay rate	
τ_{learn}	0.2	Reward location learn rate	
ν	2	Place cells smoothness	
N_{GP}	36	No of taxon input directions	Arbitrary, the precision can be higher than $2\pi/N_{GP}$ degrees because it is a population code.
N_{dir}	36	No of taxon output directions	See N_{GP}
η	0.1	Taxon learning rate	
α	0.8	Taxon future reward discount rate	
κ	0.5	Taxon eligibility traces decay factor	
ω	$\pi/8$	Taxon action generalization	

References

[1] Meyer J-A, Guillot A, Girard B, Khamassi M, Pirim P and Berthoz A 2005 The Psikharpax project: towards building an artificial rat *Robot. Auton. Syst.* **50** 211–23

[2] N'Guyen S, Pirim P and Meyer J-A 2010 Tactile texture discrimination in the robot-rat Psikharpax *BIOSIGNALS 2010: 3rd Int. Conf. on Bio-Inspired Systems and Signal Processing (Valencia, Spain)* pp 74–81

[3] Bernard M, N'Guyen S, Pirim P, Gas B and Meyer J-A 2010 Phonotaxis behavior in the artificial rat Psikharpax *IRIS2010: Int. Symp. on Robotics and Intelligent Sensors (Nagoya, Japan)* pp 118–22

[4] N'Guyen S 2010 Mise au point du système vibrissal du robot-rat Psikharpax et contribution à la fusion de ses capacités visuelle, auditive et tactile *PhD Thesis* Université Pierre et Marie Curie

[5] N'Guyen S, Pirim P, Meyer J-A and Girard B 2010 An integrated neuromimetic model of the saccadic eye movements for the Psikharpax robot *SAB '10: Proc. 11th Int. Conf. on Simulation of Adaptive Behavior: From Animals to Animats (Paris, France, 25–28 August 2010) (LNAI* vol 6226*)* ed S Doncieux *et al* (Berlin: Springer) pp 188–98

[6] Trullier O, Wiener S, Berthoz A and Meyer J-A 1997 Biologically-based artificial navigation systems: review and prospects *Prog. Neurobiol.* **51** 483–544

[7] Redish A D 1999 *Beyond the Cognitive Map: From Place Cells to Episodic Memory* (Cambridge, MA: MIT Press)

[8] Khamassi M 2007 Complementary roles of the rat prefrontal cortex and striatum in reward-based learning and shifting navigation strategies *PhD Thesis* Université Pierre et Marie Curie

[9] Arleo A and Rondi-Reig L 2007 Multimodal sensory integration and concurrent navigation strategies for spatial cognition in real and artificial organisms *J. Integr. Neurosci.* **6** 327–66

[10] Morris R 1984 Developments of a water-maze procedure for studying spatial learning in the rat *J. Neurosci. Methods* **11** 47–60

[11] Packard M G, Hirsh R and White N M 1989 Differential effects of fornix and caudate nucleus lesions on two radial maze tasks: evidence for multiple memory systems *J. Neurosci.* **9** 1465–72

[12] Burgess N 2008 Spatial cognition and the brain *Annal. NY Acad. Sci.* **1124** 77–97

[13] Alexander G E, DeLong M R and Strick P L 1986 Parallel organization of functionally segregated circuits linking basal ganglia and cortex *Annu. Rev. Neurosci.* **9** 357–81

[14] Mink J W 1996 The basal ganglia: focused selection and inhibition of competing motor programs *Prog. Neurobiol.* **50** 381–425

[15] Houk J C, Adams J L and Barto A G 1995 A model of how the basal ganglia generate and use neural signals that predict reinforcement *Models of Information Processing in the Basal Ganglia* ed J C Houk, J L Davis and D G Beiser (Cambridge, MA: MIT Press) pp 249–71

[16] Graybiel A M 1998 The basal ganglia and chunking of action repertoires *Neurobiol. Learn. Memory* **70** 119–36

[17] Yin H H and Knowlton B J 2006 The role of the basal ganglia in habit formation *Nature Rev. Neurosci.* **7** 464–76

[18] Devan B D and White N M 1999 Parallel information processing in the dorsal striatum: relation to hippocampal function *J. Neurosci.* **19** 2789–98

[19] Packard M G and Knowlton B J 2002 Learning and memory functions of the basal ganglia *Annu. Rev. Neurosci.* **25** 563–93

[20] Morris R G M 1981 Spatial localization does not require the presence of local cues *Learn. Motiv.* **12** 239–60

[21] O'Keefe J and Nadel L 1978 *The Hippocampus as a Cognitive Map* (Oxford: Clarendon)

[22] Yin H H and Knowlton B J 2004 Contributions of striatal subregions to place and response learning *Learn. Memory* **11** 459–63

[23] Albertin S V, Mulder A B, Tabuchi E, Zugaro M B and Wiener S I 2000 Lesions of the medial shell of the nucleus accumbens impair rats in finding larger rewards, but spare reward-seeking behavior *Behav. Brain Res.* **117** 173–83

[24] Pfeifer R, Lungarella M and Iida F 2007 Self-organization, embodiment, and biologically inspired robotics *Science* **318** 1088–93

[25] Arbib M, Metta G and van der Smagt P 2008 Neurorobotics: from vision to action *Handbook of Robotics* (Berlin: Springer) pp 1453–80

[26] Meyer J-A and Guillot A 2008 Biologically-inspired robots *Handbook of Robotics* ed B Siciliano and O Khatib (Berlin: Springer) pp 1395–422

[27] Arleo A and Gerstner W 2000 Spatial cognition and neuro-mimetic navigation: a model of hippocampal place cell activity *Biol. Cybern.* **83** 287–99

[28] Krichmar J L, Seth A K, Nitz D A, Fleischer J G and Edelman G M 2005 Spatial navigation and causal analysis in a brain-based device modeling corticalhippocampal interactions *Neuroinformatics* **3** 147–69

[29] Fleischer J G, Gally J A, Edelman G M and Krichmar J L 2007 Retrospective and prospective responses arising in a modeled hippocampus during maze navigation by a brain-based device *Proc. Natl Acad. Sci.* **104** 3556–61

[30] Barrera A and Weitzenfeld A 2008 Biologically-inspired robot spatial cognition based on rat neurophysiological studies *Auton. Robots* **25** 147–69

[31] Giovannangeli C and Gaussier P 2008 Autonomous vision-based navigation: goal-oriented action planning by transient states prediction, cognitive map building, and sensory-motor learning *Proc. Int. Conf. on Intelligent Robots and Systems* vol 1 (Berkeley, CA: University of California Press) pp 281–97

[32] Milford M and Wyeth G 2010 Persistent navigation and mapping using a biologically inspired slam system *Int. J. Robot. Res.* **29** 1131–53

[33] Dollé L, Sheynikhovich D, Girard B, Chavarriaga R and Guillot A 2010 Path planning versus cue responding: a bioinspired model of switching between navigation strategies *Biol. Cybern.* **103** 299–317

[34] Pearce J M, Roberts A D and Good M 1998 Hippocampal lesions disrupt navigation based on cognitive maps but not heading vectors *Nature* **396** 75–7

[35] Ragozzino M E, Detrick S and Kesner R P 1999 Involvement of the prelimbic-infralimbic areas of the rodent prefrontal cortex in behavioral flexibility for place and response learning *J. Neurosci.* **19** 4585–94

[36] Birrell J M and Brown V J 2000 Medial frontal cortex mediates perceptual attentional set shifting in the rat *J. Neurosci.* **20** 4320–4

[37] Killcross S and Coutureau E 2003 Coordination of actions and habits in the medial prefrontal cortex of rats *Cereb. Cortex* **13** 400–8

[38] Ujfalussy B, Erős P, Somogyvári Z and Kiss T 2008 Episodes in space: a modeling study of hippocampal place representation *SAB '08: Proc. 10th Int. Conf. on Simulation of Adaptive Behavior: From Animals to Animats (Osaka, Japan, 7–12 July 2008) (LNAI* vol 5040*)* ed M Asada *et al* (Berlin: Springer) pp 123–36

[39] Block M T 1969 A note on refraction and image formation of rats eye *Vis. Res.* **9** 705–11

[40] Parker A J 2007 Binocular depth perception and the cerebral cortex *Nature Rev. Neurosci.* **8** 379–91

[41] Dollé L, Sheynikhovich D, Girard B, Ujfalussy B, Chavariagga R and Guillot A 2010 Analyzing interactions between cue-guided and place-based navigation with a computational model of action selection: influence of sensory cues and training *SAB '10: Proc. 11th Int. Conf. on Simulation of Adaptive Behavior: From Animals to Animats (Paris, France, 25–28 August 2010) (LNAI* vol 6226*)* ed S Doncieux *et al* (Berlin: Springer) pp 335–46

[42] Pearson K 1901 On lines and planes of closest fit to systems of points in space *Phil. Mag.* **2** 559–72

[43] Kohonen T, Schroeder M R and Huang T S 2001 *Self-Organizing Maps* (Secaucus, NJ: Springer)

[44] Fritzke B 1995 A growing neural gas network learns topologies *Advances in Neural Information Processing Systems 7* (Cambridge, MA: MIT Press) pp 625–32

[45] Bishop C M 2007 *Pattern Recognition and Machine Learning* (Berlin: Springer)

[46] Schölkopf B, Smola A and Muller K-R 1998 Nonlinear component analysis as a kernel eigenvalue problem *Neural Comput.* **10** 1299–319

[47] Kim K I, Franz M O and Schölkopf B 2003 Kernel Hebbian algorithm for iterative kernel principal component analysis *Technical Report* Max Planck Institute for Biological Cybernetics

[48] Hasselmo M E 2005 A model of prefrontal cortical mechanisms for goal-directed behavior *J. Cognitive Neurosci.* **17** 1115–29

[49] Martinet L-E, Sheynikhovich D, Benchenane K and Arleo A 2011 Spatial learning and action planning in a prefrontal cortical network model *PLoS Comput. Biol.* **7** e1002045

[50] Banquet J P, Gaussier P, Quoy M, Revel A and Burnod Y 2005 A hierarchy of associations in hippocampo-cortical systems: cognitive maps and navigation strategies *Neural Comput.* **17** 1339–84

[51] Cuperlier N, Quoy M and Gaussier P 2007 Neurobiologically inspired mobile robot navigation and planning *Front. Neurorobot.* **1** 3

[52] Floyd R W 1962 Algorithm 97: shortest path *Commun. ACM* **5** 345

[53] Sutton R S and Barto A G 1998 *Reinforcement Learning: An Introduction* (Cambridge, MA: MIT Press)

[54] Khamassi M, Martinet L E and Guillot A 2006 Combining self-organizing maps with mixture of experts: application to an actor-critic model of reinforcement learning in the basal ganglia *SAB '06: Proc. 9th Int. Conf. on Simulation of Adaptive Behavior: From Animals to Animats (Rome, Italy, 25–29 September 2006) (LNAI* vol 4095*)* ed S Nolfi *et al* (Berlin: Springer) pp 394–405

[55] Packard M and McGaugh J 1996 Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects the expression of place and response learning *Neurobiol. Learn. Memory* **65** 65–72

[56] Save E and Poucet B 2000 Hippocampal–parietal cortical interactions in spatial cognition *Hippocampus* **10** 491–9

[57] Eilam D and Golani I 1989 Home base behavior of rats (*rattus norvegicus*) exploring a novel environment *Behav. Brain Res.* **34** 199–211

[58] Doya K 2000 Complementary roles of basal ganglia and cerebellum in learning and motor control *Curr. Opin. Neurobiol.* **10** 732–9

[59] Samejima K, Ueda Y, Doya K and Kimura M 2005 Representation of action-specific reward values in the striatum *Science* **310** 1337–40

[60] Roesch M R, Calu D J and Schoenbaum G 2007 Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards *Nature Neurosci.* **10** 1615–24

[61] Matsumoto M, Matsumoto K, Abe H and Tanaka K 2007 Medial prefrontal cell activity signaling prediction errors of action values *Nature Neurosci.* **10** 647–56

[62] Chavarriaga R, Strösslin T, Sheynikhovich D and Gerstner W 2005 A computational model of parallel navigation systems in rodents *Neuroinformatics* **3** 223–41

[63] Daw N D, O'Doherty J P, Dayan P, Seymour B and Dolan R J 2006 Cortical substrates for exploratory decisions in humans *Nature* **441** 876–9

[64] Watkins C J C H and Dayan P 1992 Technical note: Q-learning *Mach. Learn.* **8** 279–92

[65] Miller E K and Cohen J D 2001 An integrative theory of prefrontal cortex function *Annu. Rev. Neurosci.* **24** 167–202

[66] Peyrache A, Khamassi M, Benchenane K, Wiener S I and Battaglia F P 2009 Replay of rule-learning related neural patterns in the prefrontal cortex during sleep *Nature Neurosci.* **12** 919–26
[67] Uylings H B, Groenewegen H J and Kolb B 2003 Do rats have a prefrontal cortex? *Behav. Brain Res.* **146** 3–17
[68] Granon S and Poucet B 2000 Involvement of the rat prefrontal cortex in cognitive functions: a central role for the prelimbic area *Psychobiology* **28** 229–37
[69] Baeg E H, Kim Y B, Huh K, Mook-Jung I, Kim H T and Jung M W 2003 Dynamics of population code for working memory in the prefrontal cortex *Neuron* **40** 177–88
[70] Hok V, Save E, Lenck-Santini P P and Poucet B 2005 Coding for spatial goals in the prelimbic/infralimbic area of the rat frontal cortex *Proc. Natl Acad Sci. USA* **102** 4602–7
[71] Mulder A B, Nordquist R E, Orgut O and Pennartz C M 2003 Learning-related changes in response patterns of prefrontal neurons during instrumental conditioning *Behav. Brain Res.* **146** 77–88
[72] Kargo W J, Szatmary B and Nitz D A 2007 Adaptation of prefrontal cortical firing patterns and their fidelity to changes in action-reward contingencies *J. Neurosci.* **27** 3548–59
[73] Daw N D, Niv Y and Dayan P 2005 Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control *Nature Neurosci.* **8** 1704–11
[74] Ostlund S B and Balleine B W 2005 Lesions of medial prefrontal cortex disrupt the acquisition but not the expression of goal-directed learning *J. Neurosci.* **25** 7763–70
[75] Johnson A and Redish A D 2007 Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point *J. Neurosci.* **27** 12176–89
[76] Johnson A, van der Meer M A A and Redish A D 2007 Integrating hippocampus and striatum in decision-making *Curr. Opin. Neurobiol.* **17** 692–7
[77] Meer M A A van der and Redish A D 2011 Theta phase precession in rat ventral striatum links place and reward information *J. Neurosci.* **31** 2843–54
[78] van der Meer M A A and Redish A D 2011 Ventral striatum: a critical look at models o learning and evaluation *Curr. Opin. Neurobiol.* **21** 387–92
[79] Salazar R F, White W, Lacroix L, Feldon J and White I M 2004 NMDA lesions in the medial prefrontal cortex impair the ability to inhibit responses during reversal of a simple spatial discrimination *Behav. Brain Res.* **152** 413–24
[80] Naneix F, Marchand A R, DiScala G, Pape J R and Coutureau E 2009 A role of the medial prefrontal cortex dopaminergic innervation in instrumental conditioning *J. Neurosci.* **29** 6599–606
[81] Battaglia F P, Peyrache A, Khamassi M and Wiener S I 2008 Spatial decisions and neuronal activity in hippocampal projection zones in prefrontal cortex and striatum *Hippocampal Place Fields: Relevance to Learning and Memory* ed S Mizumori (Oxford: Oxford University Press) chapter 18, pp 289–311
[82] Rich E L and Shapiro M 2009 Rat prefrontal cortical neurons selectively code strategy switches *J. Neurosci.* **29** 7208–19
[83] Bonasso P and Dean T 1997 A retrospective of the AAAI robot competitions *AI Mag.* **18** 11–23
[84] Gat E 1998 On three-layer architectures *Artificial Intelligence and Mobile Robots: Case Studies of Successful Robot Systems* ed D Kortenkamp, R P Bonnasso and R Murphy (Cambridge, MA: MIT Press) pp 195–210
[85] Kortenkamp D and Simmons R 2008 Robotic systems architectures and programming *Handbook of Robotics* ed B Siciliano and O Khatib (Berlin: Springer) pp 187–206
[86] Minguez J, Lamiraux F and Laumond J P 2008 Motion planning and obstacle avoidance *Handbook of Robotics* ed B Siciliano and O Khatib (Berlin: Springer) pp 827–52
[87] Dickinson A 1985 Actions and habits: The development of behavioural autonomy *Phil. Trans. R. Soc.* B **308** 67–78
[88] Keramati M, Dezfouli A and Piray P 2011 Speed/accuracy trade-off between the habitual and goal-directed processes *PLoS Comput. Biol.* **7** 1–25
[89] Comon P 1994 Independent component analysis, a new concept? *Signal Process.* **36** 287–314
[90] MacQueen J B 1967 Some methods for classification and analysis of multivariate observations *Proc. 5th Berkeley Symp. on Mathematical Statistics and Probability* vol 1 ed L M Le Cam and J Neyman (Berkeley, CA: University of California Press) pp 281–97
[91] Zito T, Wilbert N, Wiskott L and Berkes P 2009 Modular toolkit for data processing (MDP): a python data processing framework *Front. Neuroinform.* **2** 8
[92] Mahalanobis P C 1936 On the generalised distance in statistics *Proc. National Institute of Science (India)* vol 2 pp 49–55

4.2 Habit learning in a humanoid robot

4.2.1 Renaudo, Girard, Chatila, Khamassi (2014)

Design of a Control Architecture for Habit Learning in Robots

Erwan Renaudo[1,2], Benoît Girard[1,2], Raja Chatila[1,2], and Mehdi Khamassi[1,2]

[1] Sorbonne Universités, UPMC Univ Paris 06, UMR 7222, Institut des Systèmes Intelligents et de Robotique, F-75005, Paris, France
[2] CNRS, UMR 7222, Institut des Systèmes Intelligents et de Robotique, F-75005, Paris, France

Abstract. Researches in psychology and neuroscience have identified multiple decision systems in mammals, enabling control of behavior to shift with training and familiarity of the environment from a goal-directed system to a habitual system. The former relies on the explicit estimation of future consequences of actions through planning towards a particular goal, which makes decision time longer but produces rapid adaptation to changes in the environment. The latter learns to associate values to particular stimulus-response associations, leading to quick reactive decision-making but slow relearning in response to environmental changes. Computational neuroscience models have formalized this as a coordination of model-based and model-free reinforcement learning. From this inspiration we hypothesize that it could enable robots to learn habits, detect when these habits are appropriate and thus avoid long and costly computations of the planning system. We illustrate this in a simple repetitive cube-pushing task on a conveyor belt, where a speed-accuracy trade-off is required. We show that the two systems have complementary advantages in these tasks, which can be combined for performance improvement.

1 Keywords

Adaptive Behaviour • Habit Learning • Reinforcement Learning • Robotic Architecture

2 Introduction

Researches in the field of instrumental conditioning in psychology have shown that rodents learning to press a lever in order to get food progressively shift from a goal-directed decision system to a habitual system [7,8]. After moderate training, devaluation of the outcome (e.g. pairing it with illness) leads the animal to quickly stop pressing the lever. In contrast, after extensive training the animal perseveres with pressing the lever even after outcome devaluation - hence "habit" [1,21]. This has been hypothesized to enable the animal to avoid slow and costly decision-making through planning by shifting to reactive decision-making when

the stability of the environment makes habits reliable, a capacity which is shared with humans and other mammals [2].

In contrast, current robots are still rarely equipped with efficient online learning abilities and mostly rely on a single planning decision-making system, thus not providing alternative solutions to motion planning in situations where such strategy is limited [17]. Indeed the planning strategy can be approximative when coping with uncertainties, *e.g.* when there is perceptual aliasing [4], and can also require high computational costs and long time to propagate possible trajectories through internal representations [10]. We have previously shown that taking inspiration from the way rodents shift between different navigation strategies – a capacity which has been shown to be analogous to the shifts between goal-directed and habitual decision systems [14] – can be applied to a robotic platform to enable to automatically exploit the advantages of each strategy [4,3]. However, these experiments only involved navigation behaviors from one location to another. To our knowledge, no application has yet been made of the coordination of goal-directed and habitual systems to robotic tasks.

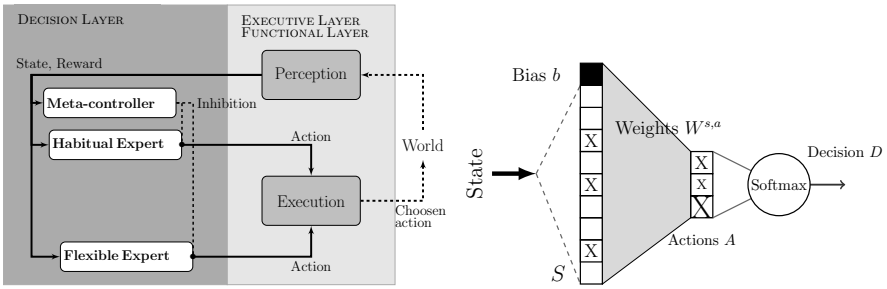
In this work, we illustrate the application of a decision architecture combining a goal-directed expert with a habitual one to a simple task where a simulated robot have to learn to repeat the less costly sequence of actions to push a series of cubes arriving in front of him on a conveyor belt. We build our algorithm on computational neuroscience models which have shown that combining model-based and model-free reinforcement learning can accurately reproduce properties of the competition between goal-directed and habitual systems [5,13,9]. In these models, the goal-directed system is modelled with model-based reinforcement learning in the sense that the system plans sequences of actions towards a particular goal by using the transition and reward functions. In parallel, the model-free reinforcement learning progressively learns by trial-and-error the Q-values associated to different state-action couples. The criterion for switching from one system to the other is based on the measure of uncertainty in the model-free system: the less variance there is in the Q-values, the more reliable the model-free habitual system is considered and the more likely it will control the behavior of the simulated agents.

In contrast to these previous computational neuroscience models, we do not a priori give the transition and reward functions (i.e. the considered model of the task) to the algorithm but rather make it learn it automatically by observing experienced transitions and rewards. Moreover, we arbitrate without bias between systems, as the selection of each one is random and equiprobable. The task that we simulate requiring a certain balance between speed and accuracy so as not to skip some cubes coming on the conveyor belt, our simulations show that the two systems have complementary advantages that can be combined for a highest performance. In a first series of simulations where the systems are controlling individually the agent, we characterize their performances in a constant belt velocity and constant distance between cubes setup and when the belt velocity is changed during the simulation. We show that each system is performing differently to these conditions as the model-free is more efficient than the model-

based to exploit the stability in the environment, but the model-based adapts quickly to condition changes in the environment. We then show how combining the two systems and switching control among them, even with a basic rule, can improve the robot policy and gives it the ability to perform well both in a stable environment and during transitional phases to another stable setup, with the same architecture.

3 Materials and Methods

3.1 Global Architecture



(a) The inner structure of the Decision Layer and its connection with the Executive and Functional Layers.

(b) The Habitual Expert neural network. S is the state mapped on input, A the action vector, $W^{s,a}$ the weights. Black X represents activity in the network, sized to activity.

Fig. 1: Robotic organisation of modules and Habitual Expert structure

Our Decision Layer [10] consists of two Experts that learn a policy and a Meta-Controller that supervises the Experts' performance (Fig. 1a). The Flexible Expert is a Model-Based Reinforcement Learning agent and the Habitual Expert is a Model Free reinforcement learning agent [19].

These modules receive the current State $S \in \mathcal{S}$ from a Perception Module and choose their actions in a set $\mathcal{A}$. Each Expert decides, from the current State and their knowledge, which action to take. In parallel, the Meta-Controller decides which Expert is the most efficient in the current State and allows it to send its action choice to be executed.

3.2 Habitual Expert

The Habitual Expert (MF) is implemented as a 1-layer neural network. It learns directly the relevant state-action policy without an internal representation of transitions between states of the world (hence the term Model Free). Propagating

the values from input S (and bias b) to action output A is computationally cheap, but learning the whole policy is long: only the experienced state-action value is updated. Learning a new policy to adapt to a new environment configuration is longer than just learning the first policy so this expert is reluctant to changes.

$$A_t(i) = W_t \cdot S_t + b_t(i) \ . \tag{1}$$

The connection weights W_t that learn the State-Action association are updated according to a Qlearning rule [20]: the connection between input neurons encoding the previous State S_{t-1} and the output neuron of the action done is modified with an amount depending on the reward R obtained.

$$\delta = R_t(s_{t-1}, a_{t-1}) + \gamma_{MF} \cdot max_a \left(W_{t-1}(a) \cdot S_t\right) - \left(W_{t-1}(a_{t-1}) \cdot S_{t-1}\right) \ . \tag{2}$$

$$W_t(a_{t-1}) = W_{t-1}(a_{t-1}) + \alpha_{MF} \cdot \delta \ . \tag{3}$$

α_{MF} : learning rate, γ_{MF} : decay factor.

Each action activity is interpreted as the probability $P(A(i))$ of taking action $A(i)$, using a Softmax rule (4). The decision is taken stochastically in the resulting distribution (τ_{MF} : temperature.).

$$P_t(A_t(i)) = \frac{\exp\left(\frac{A_t(i)}{\tau_{MF}}\right)}{\sum_j \exp\left(\frac{A_t(j)}{\tau_{MF}}\right)} \ . \tag{4}$$

3.3 Flexible Expert

The Flexible Expert (MB) is a Model Based Reinforcement Learning agent. It learns a model of the *Transition* and *Reward* functions of the task. The former is a *cyclic graph* of States connected by Actions, the latter a *table* of (State, Action) and Reward association. Decisions are taken based on these representations of the world. As the problem topology is modeled, a change experienced in the environment (ie. a transition leads to a new state) can quickly be handled by updating the model, allowing the next decision to be adapted to the changes.

The Reward function is learned from the experienced transition and is directly the instant reward obtained $R_t(S, A) = R_t$. The Transition function is progressively learned according to (5). The probability T of experienced transition $S \xrightarrow{A} S'$ is updated at learning rate α_{MB}.

$$T_t(S, A, S') = T_{t-1}(S, A, S') + \alpha_{MB} \cdot \left(1 - T_{t-1}(S, A, S')\right) \ . \tag{5}$$

Planning with the models consists in computing the Quality Q(S,A) of performing action A in the given state S. It is done iteratively by propagating the known rewards and refining the estimated Quality value according to the Transition function until convergence (γ_{MB} : decay factor):

$$Q_t(s,a) = max\left(R_t(s,a), (\gamma_{MB} \cdot \sum_{s'} T_{t-1}(s,a,s') \cdot \max_{a'} Q_t(s',a'))\right) \quad . \tag{6}$$

The Decision is also taken with the *softmax rule* (cf. Eq. (4)).

The drawback of such a method comes from the increasing size of the transition model. Planning becomes more and more time consuming and the Expert is less and less reactive. As the environment evolves, even in a predictable way, the action decided from the perceived State at S_{t-1} may be irrelevant when acting in State S_t. To improve the Flexible Expert performance and keep a manageable model while dimensionality increases, the following features are implemented :

1. Planning in the graph is bounded in time : if planning is longer than a certain time chosen in agreement with the task dynamics, the computation of Quality is stopped and the approximated values are used for decision, as it is more important to be reactive enough than having accurate values in this task.
2. As the best policy is learnt, fewer and fewer states are visited, producing a peaked distribution V_S of states visits. The allocated computation time being limited, planning should only consider the most visited states. These states are hypothetized to be the most interesting for the Expert, as the policy focuses on a subset of all experienced states. A subgraph of the N most visited states is extracted to have their Q-values computed as a priority. To have a relevant value for N, we compute the entropy of the V_S distribution, getting a measure of the model organisation :

$$H(V_S) = -\sum_{i \in S} P(i) \cdot log_2\left(P(i)\right) \text{ with } P(i) = \frac{Card(V_{S_i})}{Card(V_S)} \quad . \tag{7}$$

We compare this measure to the maximal entropy of the model, deducing a ratio R_c of the compressibility of the State distribution representation :

$$R_c = \frac{H(V_S)}{H_{max}(V_S)} \text{ with } H_{max}(V_S) = log_2|\mathcal{S}| \quad . \tag{8}$$

This method guides the planification to the most visited states. The drawback is that it may erroneously limit the use of the model during early states of learning where the number of states is small but the distribution already presents a contrasted shape. In this case, planning in the full graph is still possible at reasonable cost. To avoid this behaviour, R_c is transformed into a Ratio R_n - depending on the known number of nodes - given the following function (9).

$$R_n = (1-\omega) + \omega \cdot R_c \text{ with weight } \omega = \frac{1}{1+e^{-\sigma|\mathcal{S}|}} \quad . \tag{9}$$

The final number of states to plan on is a proportion of the number of known states $|\mathcal{S}|$:

$$N = R_n \cdot |\mathcal{S}| \quad . \tag{10}$$

3.4 Meta-Controller

The Meta-Controller gives the control to one of the Experts given a criterion. It allows only one of the Experts to send its decision to the Execution Layer. It also sends back the decision to both Experts, such that they can update their knowledge about its relevance in the current state according to the feedback, and cooperate in learning the best policy. The criterion considered in this work is an equiprobable random selection of each Expert, as a proof of concept of the interest of combining the two.

4 Results

4.1 Experiment Description

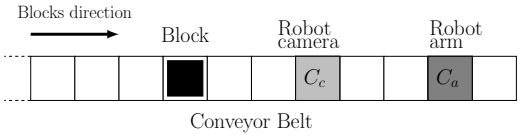


Fig. 2: The experimental setup : a discrete conveyor belt is carrying blocks in front of the robot. The robot's camera points at space C_c and its arm can reach the space in C_a. Blocks are going from left to right such that a block can be first seen and then touched.

We evaluated our Architecture performance in the simulation of a simple task of block pushing. The system has been implemented using the ROS middleware [18]. Our simulated robot is facing a conveyor belt on which are placed blocks. These blocks are characterized by their velocity (BS) and the distance between two blocks (inter-block distance, or IBD). These simulation parameters may be constant or evolve during the experiment, leading to four different cases. In this work, we focused on :

1. Regular case : inter-block distance is constant, speed of blocks is constant.
2. Speed Shift case : IBD is constant, BS changes during experiment.

In our setup, acting is required to update the perception (see Sect. 4.2 for Perception Module description). The robot has three available actions :

1. **Do nothing (DN) :** this action doesn't modify the environment nor bring perceptual information. It is a waiting action with no cost ($R_t = 0$) when executed.
2. **Look Cam (LC) :** this action doesn't modify the environment but updates the view modality about the presence of a block in C_c. It has a cost of $R_t = -0.03$.
3. **Push Arm (PA) :** this action can modify the environment : if a block is in C_a and PA is done, it is removed from the belt. The contact modality is updated about the perceived block. The action costs $R_t = -0.03$ but brings a positive reward when a block is pushed for a final reward of $R_t = 0.97$).

4.2 Perception Module

The Perception Module (Fig. 3) transforms Perceptions into States. Our simulated robot is equipped with a visual block detector simulated camera (signal p^{bs}) and a tactile binary sensor on its arm (signal p^{bt}). When the corresponding action is selected (LookCam for p^{bs} and PushArm for (p^{bt}), these informations update memories where each element is one step further in the past. Each modality has its own memory where older block perceptions are recorded. Memories (M_{bs}, M_{bt}) have a finite length (8 elements) such that the system only considers closest perceptions. Each configuration of both memories defines a unique State, used by the Experts.

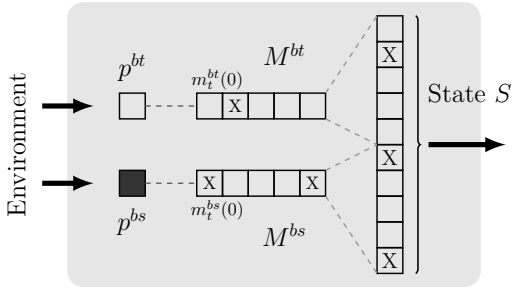


Fig. 3: Perception module for our task. The X corresponds to a memorized block. The system visually perceives a block and updates the corresponding memory.

The perceptive input are binary and their perceptions are determined according to (11).

$$p^{bt} = C_a \cdot \text{PushArm}, \, p^{bs} = C_c \cdot \text{LookCam} \ . \tag{11}$$

Memories can evolve in two ways : if enough time has elapsed (State Max Duration = 0.1s here) or when a new perception brings information. In both cases, all elements from the memory are shifted to the next timestep. Information that exceeds memory length is forgotten. The first memory element is populated with the relevant perceptive data.

$$\begin{cases} m_t^{bs,bt}(i) = m_{t-1}^{bs,bt}(i-1) \ \forall i \in |M| \ . \\ m_t^{bs,bt}(0) = p^{bs,bt} \ . \end{cases} \tag{12}$$

4.3 Parameters search

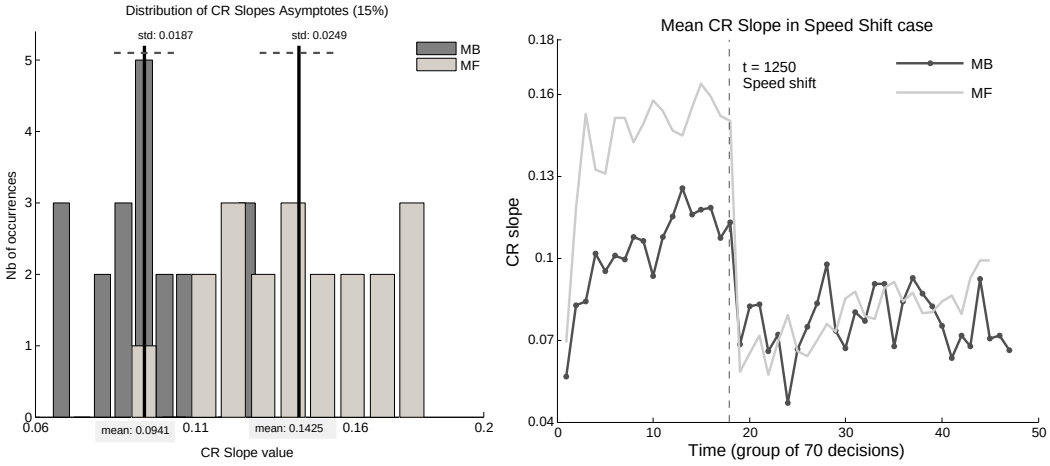
In the following, we consider : $BS = 8$ spaces/s (optimal policy : DN-DN-DN-DN-PA), $IBD = 4$ spaces/block.

We first searched for the best parametrization for both Experts controlling individually the robot, in the Regular case. An Expert is performant if

it maximises the obtained Cumulative Reward (CR) and minimizes the standard deviation of CR over runs. We tested for each Expert the combination of 3 to 5 values (for MB and MF : $\alpha_{MB,MF} \in \{0.01, 0.05, 0.1, 0.5, 0.9\}, \gamma_{MB,MF} \in \{0.5, 0.98, 0.9999\}, \tau_{MB,MF} \in \{0.05, 0.1, 0.5, 0.9\}$ plus $\tau_{MB} = 0.01$) :

For the best solutions, we favor the most rewarding in mean and then the less varying. We choose $\alpha_{MB} = 0.5, \gamma_{MB} = 0.5, \tau_{MB} = 0.1$ and $\alpha_{MF} = 0.1, \gamma_{MF} = 0.9999, \tau_{MF} = 0.05$.

4.4 Individual experts performances



(a) Regular case (20 runs) (b) Speed shift case (10 runs)

Fig. 4: Policy evaluations. (4a) histogram of approximated slopes. Solid lines are means of CR slope for each Expert, dashed lines the standard deviation (4b) Mean CR slope over time. The slope is approximated every 70 decisions.

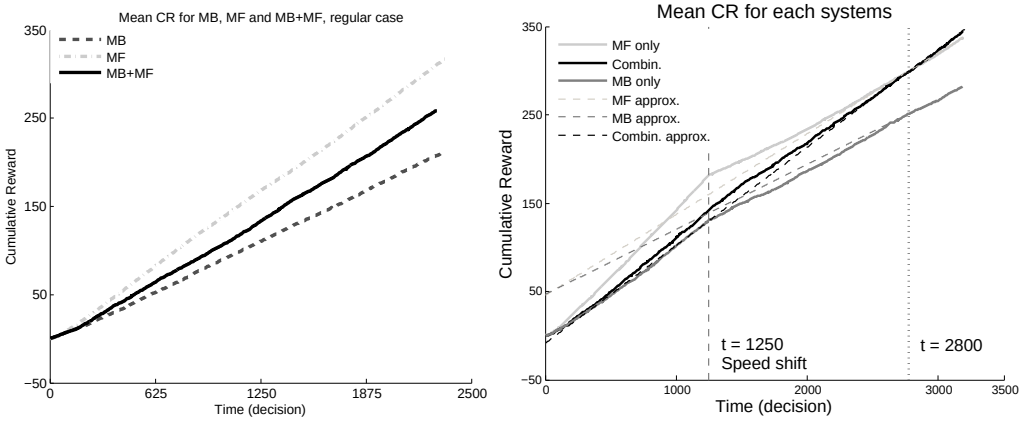
Each Expert is tested individually in the Regular and Speed Shift cases (simulation parameters : see Sect. 4.3 ; in Speed Shift case, we change BS to 13.2 spaces/s at 1250 decisions, which correspond to an optimal DN-DN-PA policy). The Cumulative Reward is linearly approximated from its 15% last values to evaluate the discovered policy. For the Speed shift case, we also approximate CR before the speed shift (on the same duration) and compare it to the first approximation to evaluate the sensitivity of the policy to speed shift. The slope of these approximations measures the quality of the policy as it depends on the obtained rewards. From figure 4a, we observe that the MF discovers and follows better policies in mean but tends to be more exploring than the MB with a larger deviation in slopes values. In the Regular case, the MF is more relevant than the MB to obtain the best performance as possible. This is due to the low

cost and high precision of the Q-values acquired by the MF. In contrast, the MB learns a model of the task whose number of states rapidly grows, making the planning process slow, costly and relying on approximations of action values.

In the Speed Shift case, we observe a break in Cumulative reward for the MF. Figure 4b shows that in mean, the MB performance is less sensitive to the change than the MF : the environmental change induces a loss that is more than twice higher in the MF than in the MB. After the shift and until the end of simulation, both MB and MF are performing similarly. This behavior can be explained by the long time required by the MF to relearn the Q-values of a new efficient action sequence, what doesn't happen in the given time. In contrast, a single exposure of the MB to the new sequence of events imposed by the speed shift enables it to change its model of the task and thus to plan a new sequence of actions, but it still suffer from the approximation of action values to find a better policy.

Both Experts exhibit a complementary role : while the MF is best suited to optimize the policy in stable conditions, the MB can better handle transient phases following environmental changes.

4.5 Combination of Experts



(a) Regular case (20 runs) (b) Speed shift case (10 runs)

Fig. 5: Mean Cumulative Reward obtained from individual Experts and their Combination (solid line).

The whole architecture (MB+MF, supervised by Meta-Controller) is then tested on both cases, with the same setup. In the Regular case, figure 5a shows that the strategy of selecting stochastically each Expert improves the mean performance of the robot compared to using only the MB. On the other hand, as the

Experts are chosen randomly, the robot is not relying only on the MF, which is the most efficient strategy in the Regular case. This explains that the MB+MF performance is still worse than the MF only. In the Speed Shift case, figure 5b shows that the change in the environment doesn't affect significantly the robot's performance, as it benefits from the MB ability to quickly replan an adapted policy. The Combination of Experts robustness to changes compensates for the advantage gained by the MF before the shift. At the end of the simulation, the MF hasn't found a policy that is at least as good as before the shift (though the task allows a higher rate of reward, as there are more blocks to push on).

5 Discussion

This work presented the decision layer of a robotic control architecture able to learn habits, taking inspiration from computational neuroscience models [13] and multiple reinforcement learning systems applied to navigation [9,4]. The model has two different Experts, or strategies, one habitual – that learns State-Action association, and make quick decisions, but slowly adapts its policy when the environment changes – and one flexible – that maintains a representation of the environment and the task, can adapt quickly but is slow in deciding as it evaluates action outcomes. These strategies are selected depending on an arbitration criterion by a Meta-Controller. The criterion used in this work is an random equiprobable selection of Experts, as a proof-of-concept of the interest of combining the two.

We first highlighted each Expert properties in a Regular and a Speed Shift cases. We showed that, as expected, the Habitual Expert learns better policies than the Flexible Expert when the environment is stable, but a transient phase like a shift in belt speed, making the policy less appropriate, will result in a long lower performance period. The learnt Flexible Expert policies are less performant in mean as the computation time constraint and the focused planning lead to less precise Q-values when the number of states grows, and a sub-optimal policy. On the other hand, updating its model allows the Flexible Expert to be less affected by the speed shift. We then showed that a random selection of each Experts is able to benefit from the shift robustness of the Flexible Expert while the rewarding policies from the Habitual Expert improve the global performance of the robot.

These results show that the multiple reinforcement learning systems approach is relevant to handle complex environments that can evolve during the robot operating period. The combination at same level of MB and MF can improve the robot autonomy provided that the MB is designed to be reactive to the environment dynamics. It has to be able to decide in parallel with the MF in order to remain useful for control. Indeed, a classical task in neuroscience is usually modelled by a Markov Decision Process with few states and actions (e.g. intrumental task of pressing a lever and entering a magazine to get food [6,13,16]) but the

dimensionality is much higher when reinforcement learning is applied to robotics [12,15], as states are discretized from robot's perceptions. In our task, we end up with several hundreds of states and we need to bound the computation time and focus planning on the hypothesized most interesting states. This justifies the need for a mechanism that manage the known states within the MB system, a proposition which has recently been applied to Computational Neuroscience models [11]. In this work, we first tested an exponential forgetting mechanism on the transition model to remove unvisited paths. As this mechanism doesn't strongly affect performance and planning time is still increasing with the growth of states, we switched to the time constraint and focused planning mechanism (described in Sect. 3.3). The increase in performance suggests that planning with a complex model requires a budgeted approach that only considers the relevant sub-model, instead of pruning parts of the model related to irrelevant old experiences.

This work also generalizes the concepts from [4] for the control of robots. The multiple reinforcement learning systems approach can be applied not only for navigation but also on a wider variety of tasks, provided that the robot is able to perform the relevant actions. As our system can rely on the Habitual Expert, our architecture can benefit from its properties of being quick to decide the next action when the task is stationary. On the other hand, our Flexible Expert can be compared to the robotic decision-making systems, that are based mostly on planning algorithms that use a representation of the world [17]. The latter usually rely on a provided representation whereas our Flexible Expert learns its model and updates it according to changes in the task. This enhances the behavioral adaptability of the robot in non-stationary environments. We need to further investigate the arbitration criterion between Experts to get the optimal alternations and benefit from the whole architecture.

Acknowledgements

This work has been funded by a DGA (French National Defence Agency) scholarship (ER), by the Project HABOT from Ville de Paris and by French Agence Nationale de la Recherche ROBOERGOSUM project under reference ANR-12-CORD-0030.

References

1. B. W. Balleine and A. Dickinson. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, 37:407–419, 1998.
2. B. W. Balleine and J. P. O'Doherty. Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, 35:48–69, 2010.
3. K. Caluwaerts, A. Favre-Félix, M. Staffa, S. N'Guyen, C. Grand, B. Girard, and M. Khamassi. Neuro-inspired navigation strategies shifting for robots: Integration

of a multiple landmark taxon strategy. In T.J. et al. Prescott, editor, *Living Machines 2012, LNAI*, volume 7375/2012, pages 62–73. 2012.
4. K. Caluwaerts, M. Staffa, S. N'Guyen, C. Grand, L. Dollé, A. Favre-Félix, B. Girard, and M. Khamassi. A biologically inspired meta-control navigation system for the psikharpax rat robot. *Bioinspiration & Biomimetics*, 2012.
5. N.D. Daw, Y. Niv, and P. Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12):1704–1711, 2005.
6. Amir Dezfouli and Bernard W. Balleine. Habits, action sequences and reinforcement learning. *European Journal of Neuroscience*, 35(7):1036–1051, 2012.
7. A. Dickinson. *Contemporary animal learning theory.* Cambridge: Cambridge University Press, 1980.
8. A. Dickinson. Actions and habits: The development of behavioral autonomy. *Philosophical Transactions of the Royal Society (London)*, 308:67 – 78, 1985 1985.
9. L. Dollé, D. Sheynikhovich, B. Girard, R. Chavarriaga, and A. Guillot. Path planning versus cue responding: a bioinspired model of switching between navigation strategies. *Biological Cybernetics*, 103(4):299–317, 2010.
10. E. Gat. On three-layer architectures. In *Artificial Intelligence and Mobile Robots.* MIT Press, 1998.
11. Q.J. Huys, N. Eshel, E. O'Nions, L. Sheridan, P. Dayan, and J.P. Roiser. Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Computational Biology*, 8(3):e1002410, 2012.
12. L.P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: a survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.
13. M. Keramati, A. Dezfouli, and P. Piray. Speed/accuracy trade-off between the habitual and goal-directed processes. *PLoS Computational Biology*, 7(5):1–25, 2011.
14. M. Khamassi and M.D. Humphries. Integrating cortico-limbic-basal ganglia architectures for learning model-based and model-free navigation strategies. *Frontiers in Behavioral Neuroscience*, 6:79, 2012.
15. J Kober, D. Bagnell, and J. Peters. Reinforcement learning in robotics: A survey. (11):1238–1274, 2013.
16. F. Lesaint, O. Sigaud, S. B. Flagel, T. E. Robinson, and M. Khamassi. Modelling Individual Differences in the Form of Pavlovian Conditioned Approach Responses: A Dual Learning Systems Approach with Factored Representations. *PLoS Comput Biol*, 10(2):e1003466+, February 2014.
17. J. Minguez, F. Lamiraux, and J.P. Laumond. Motion planning and obstacle avoidance. In B. Siciliano and O. Khatib, editors, *Handbook of Robotics.*, pages 827–852. Springer-Verlag, 2008.
18. M. Quigley, K. Conley, B. P. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng. Ros: an open-source robot operating system. In *ICRA Workshop on Open Source Software*, 2009.
19. Richard S. Sutton and Andrew G. Barto. *Introduction to Reinforcement Learning.* MIT Press, Cambridge, MA, USA, 1st edition, 1998.
20. C. Watkins. *Learning from Delayed Rewards.* PhD thesis, King's College, Cambridge, UK, 1989.
21. H. H. Yin, S. B. Ostlund, and B. W. Balleine. Reward-guided learning beyond dopamine in the nucleus accumbens: the integrative functions of cortico-basal ganglia networks. *Eur J Neurosci*, 28:1437–1448, 2008.

Discussion

5

Contents

5.1 Discussion of the results

5.1.1 Discussion of the modelling results

Chapter 2 presented two computational models based on the model-based / model-free reinforcement learning framework. These models were both designed to account for behavioral data in rodents, the former during navigation tasks, the latter during a Pavlovian conditioning paradigm.

The first model has been designed with Mark D. Humphries. The aim was to show the relevance of using the model-based / model-free reinforcement learning computational framework to categorize navigation strategies in rodents and their underlying neural substrates (Khamassi and Humphries 2012). The proposed computational framework suggests that navigation strategies can be categorized as model-based or model-free, depending on the usage of information rather than on the type of information (*e.g.* cue versus place) as previous taxonomies propose. It moreover proposes that the Ventral Striatum (VS) participates to the model-building part of the involved computational processes.

The second part of the chapter presented the work of PhD student Florian Lesaint and showed that a computational model for the coordination of MB and MF RL enables to reproduce inter-individual behavioral and neurophysiological differences observed in rats called *sign-trackers* and *goal-trackers* in a Pavlovian conditioning paradigm (Lesaint et al. 2014). The simulations suggest that the behavior of both types of animals is the result of a weighted sum of MB and MF learning systems, with *sign-trackers*' behavior relying on a stronger weighting of the MF system while *goal-trackers*' behavior can be reproduced by a stronger weighting of the MB system. The model also explains why learning in *goal-trackers* has been experimentally shown to be dopamine-independent while this is not the case in *sign-trackers*.

Importantly, both models gave birth to a series of experimentally testable predictions, some having been written in a recently submitted paper (Lesaint et al. submitted). Thus a further validation of these models will require the experimental observation of these predictions.

However, these work have for the moment addressed separately the computational mechanisms underlying Pavlovian conditioning and operant behavior. In contrast, Pavlovian and instrumental conditioning are known to interact (*e.g.* see Corbit and Balleine (2011)), and some modellers have proposed that the coordination of three learning systems underlie such interactions (van der Meer et al. 2012). Thus further investigations are required to understand how multiple parallel learning processes can be combined in a single model to account for a variety of rat behavioral data in various conditions. These issues will be further addressed in the planned research project presented in section 5.2.

5.1.2 Other currently supervised modelling work

The dual reinforcement learning framework previously proposed (Daw et al. 2005, Samejima and Doya 2007) and adopted throughout this manuscript has nevertheless an important explanatory power in that it can also account for data in humans during similar conditioning paradigms than those used in rodents (Balleine and O'Doherty 2010), as well as during other paradigms.

The work of Guillaume Viejo, a first year PhD student that I cosupervise with Benoît Girard, has the goal of proposing a new computational model for the coordination of learning systems to explain human behavior in tasks involving the interaction between reinforcement learning and working memory processes. The corresponding data were recorded by Andrea Brovelli, at CNRS in Marseille, during a task where human subjects learn through trial-and-error the association between visual stimuli and finger movements. Once an association is learned, it has to be exploited during a series of repetition trials in parallel to the acquisition of other associations. Functional magnetic resonance imaging (fMRI) results during this task suggest that the dorsal striatum host complementary computations that may differentially support goal-directed and habitual processes (Brovelli et al. 2011) in the form of a dynamical interplay rather than a serial recruitment of systems.

The computational work of Collins and Frank (2012) has shown that

this type of task involves both working-memory and reinforcement learning processes, without however explicitly modelling the temporal aspect of memory manipulation. Here we develop a dual-system computational model of the two systems that can predict both performance (*i.e.*, participant choices) and modulations in reaction times during learning. One of the two systems is a model-free RL algorithm. The other one is a Bayesian working-memory algorithm which works similarly to a model-based system by searching for information in the history of previous trials. This inference process is stopped as soon as the uncertainty on the action to perform decreases below a certain threshold. An abstract has been submitted to the 2014 Computational Neuroscience meeting.

A last supervised computational modelling work addresses the question of how dopamine signals differentially impact striatal neurons with D1 and D2 receptors in a physiologically and anatomically plausible computational model of the primate basal ganglia. This work constitutes the second part of Jean Bellot's PhD thesis and is based on a biologically plausible model of primates basal ganglia, previously developed at ISIR, and which considers existing connections often neglected in the literature (Liénard and Girard 2013). Indeed, most of current basal ganglia models assume the existence of two segregated pathways : the direct pathway associated with reward and the indirect pathway associated with punishment (Frank 2005). However, if this dissociation seems to exist in mice, anatomical studies in primates revealed that these two pathways are not dissociated (Parent and Hazrati 1995a;b). While theoretical RL models are appropriate to reproduce behavior and global properties of neurophysiologicaly activity, such a neurocomputational study can contribute in capturing anatomical and physiological data at a finner scale. In particular, we are investigating the ability of the model to capture differences in reward and punishment sensitivity, with high and low-levels of dopamine, and beta oscillations observed in Parkinsonian patients.

5.1.3 Discussion of the model-based analyses' results

Chapter 3 presented work employing the model-based analysis of neurophysiological data approach. The work is presented under the form of two journal papers, one in press, the other about to be submitted, aiming at testing model predictions about hypothesized neural activities underlying behavioral adaptation, and using the computational models to more precisely measure information related to particular computational mechanisms in neural activity.

The first one started during my postdoctoral training period in the groups of Emmanuel Procyk and Peter F. Dominey at INSERM in Lyon, and showed neural substrates of adaptive regulation of reinforcement learning parameters in the prefrontal cortical network during monkey behavioral adaptation (Khamassi et al. 2014). The results show differences in activity response patterns between the Anterior Cingulate Cortex (ACC) and Lateral Prefrontal Cortex (LPFC) suggesting a role of ACC in integrating reinforcement-based information to regulate decision functions in LPFC under varying control levels, which could be interpreted in terms of

varying levels of the exploration parameter in the reinforcement learning model.

The second one presented the work of PhD student Jean Bellot and showed model-based analyses of dopamine neurons' single-unit recordings during a decision-making task in rats (Bellot et al. in preparation). The work shows that in contrast to previous reports, dopamine activity in this task only partially reflects the computation of a reward prediction error and also incorporates information about the value function, which is consistent with recent dopamine recordings challenging the classical theory (Howe et al. 2013). Moreover, the dynamics of this signal appears to be partly disconnected from the dynamics of observed behavioral adaptation, suggesting that behavior in this task is not influenced by a single learning system.

These two studies show that a computational model parametrized to fit subjects' observed behavior during a task can be used efficiently as regressors of the recorded neural activity in order to make more quantitative and computationally-grounded information measures in biological data. These two studies also contribute to confirming hypothesized dissociations between learning processes in the prefrontal cortex and the basal ganglia. However, the picture is of course incomplete. Dopamine neurons are known to also project to the prefrontal cortex (Haber et al. 2000, Lammel et al. 2008), with a hypothesized different function (Doya 2008), but whose relation to the model-based / model-free RL computational framework is not yet clear. Neither do these contributions account for different patterns of dopaminergic neurons' phasic responses (Matsumoto and Hikosaka 2009, Fiorillo 2013) and which appear to underlie different functions for learning. These issues will be further addressed in the planned research project presented in section 5.2.

5.1.4 Other currently supervised biological data analyses work

Another supervised work currently pursued and which employs model-based analyses of biological data is based on the Master research internship of Nassim Aklil. We are comparing the ability of different RL algorithms to reproduce rat behavioral data in a non-stationary multi-armed bandit task under different uncertainty levels. These data have been collected by our collaborators Alain Marchand and Etienne Coutureau at CNRS in Bordeaux, within the frame of a national ANR *Learning Under Uncertainty* project under reference ANR-11-BSV4-006, coordinated by Emmanuel Procyk at INSERM in Lyon.

One interesting aspect of this work is that classical model-free RL algorithms are compared with other algorithms non-commonly used in Neuroscience such as the upper confidence bound (UCB) method (Auer et al. 2002). This method has the advantage of proposing an optimal solution for the resolution of the exploration-exploitation trade-off. With these analyses, we are testing the hypotheses that rat behavior in this task requires a meta-learning process occuring in parallel to RL mechanisms for the dynamical regulation of the exploration parameter. We will then investigate whether the model can help us capture changes in exploration levels but not in learning performance observed in animals under injections of a do-

pamine antagonist called flupentixol (Marchand et al. 2014). This could help us further confirm the hypothesized role of dopamine in the regulation of the exploration-exploitation trade-off which we proposed in a previous computational modelling work (Humphries et al. 2012).

5.1.5 Discussion of robotic results

The work presented in Chapter 4 shows that transferring principles from biology for the coordination of model-based and model-free reinforcement learning mechanisms to robotic devices can enable flexible and adaptive behavior in autonomous robots. The models that were implemented in these robots could autonomously learn which learning system is the most appropriate and efficient to control the robot's behavior at any given moment.

Interestingly, these robotic implementations can help us learn more about properties of the tested models coming from Computational Neuroscience. First, whereas the classical hypothesis about the coordination of learning systems underlying animal behavior is that of a sequential activation (Daw et al. 2005, Yin and Knowlton 2006) – the model-based goal-directed system would control behavior during the initial phase of learning, that is during a first block of learning trials, while the model-free habitual system would take over after sufficient amount of training, that is during late blocks of learning trials –, here we found a cooperation between learning systems within single trials of the navigation task (Caluwaerts et al. 2012b). More precisely, after sufficient amount of training, the coordination system autonomously learned to trigger the model-based system to control the first actions of the behavioral sequence performed by the robot along its trajectory to the goal, while the model-free system was responsible for the last actions of the sequence, proximally to the goal. While this could result from specific properties of such a navigation task and of the used environment, these results suggest that different alternations between MB and MF control over behavior can be reached in different situations, which a supervisory system should autonomously learn based on efficient coordination mechanisms. These results are not either specific biologically-irrelevant properties of the computational mechanisms that were implemented in the robot since the computational model who inspired this robotic work has been validated on the reproduction of a set of experimental data about rat navigation behaviors (Dollé et al. 2008; 2010; submitted).

A second interesting property of these robotic results is that the coordination module autonomously learned that in some parts of the environment, neither the MB system nor the MF one were efficient to control the robot (Caluwaerts et al. 2012b). In these subparts, the robot continued to rely on its random exploration strategy even after extensive training. This was due to imperfect model of the world in the MB system in some parts of the environment, and to unreliable perception of some of the visual cues required by the MF system. This contrasts with some of the previous computational models for the coordination of MB and MF learning systems which were based on the simplified assumption that the MB system is always accurate and always has a low uncertainty (Keramati et al.

2011). Importantly, these results suggest that an efficient coordination mechanism for Robotics should be able to learn to exploit the advantages of each learning system while avoiding their drawbacks.

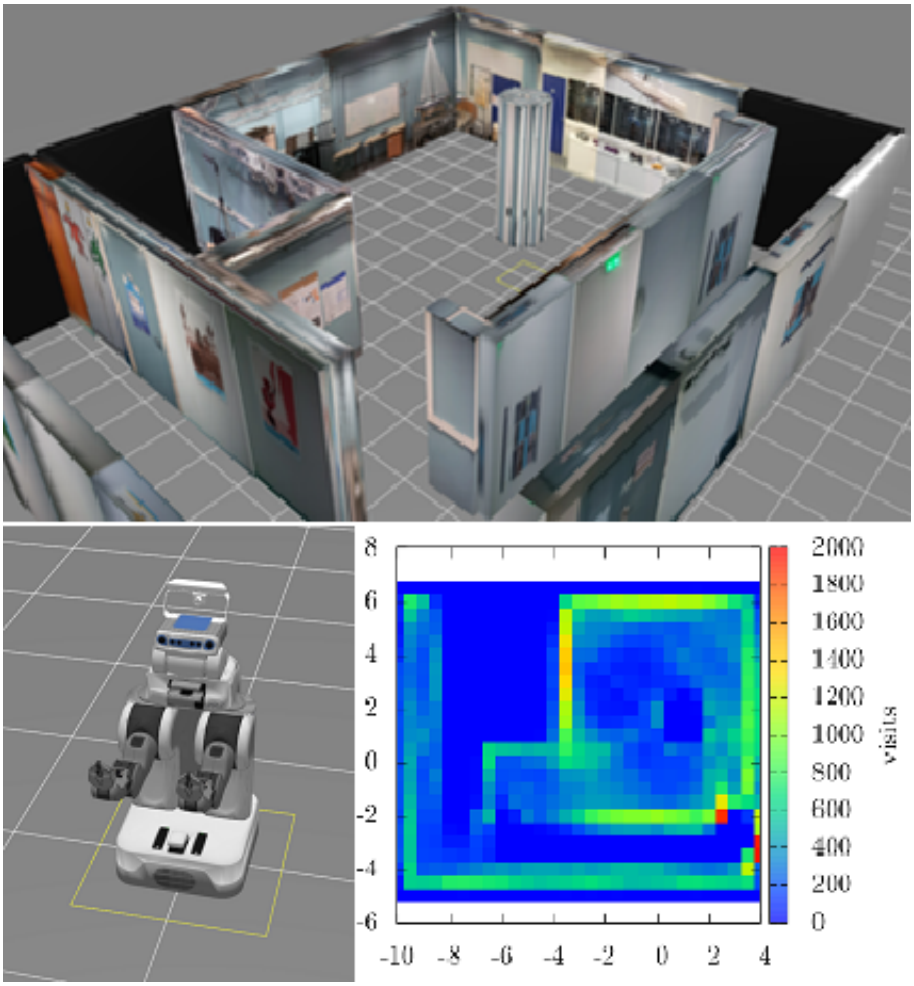


FIGURE 5.1 – *Implementations of the multiple learning systems coordination model for navigation in the PR2 Robot at ISIR*

A final interesting robotic result concerns the number of states and transitions autonomously learned within the MB system. In previous Computational Neuroscience models for the coordination of MB and MF RL systems (Daw et al. 2005, Dollé et al. 2010, Keramati et al. 2011), there is a fixed small number of states and transitions in the graph used by the MB system, which is appropriate to model the experimental tasks of interest. Here, the autonomy required for robotic implementations imposes that the MB system autonomously and incrementally builds its own model of the world. As a consequence, the number of states and transitions increases rapidly and drastically, making the planning process slower and more uncertain (Renaudo et al. 2014). This justifies the need for solutions such as a prunning mechanism within the MB system which have recently been applied to Computational Neuroscience models (Huys et al. 2012). Here we found a different solution to this problem by limiting the time for planning combined with relative weighting of states depending on the frequency with which they have been visited. It would be interesting to further investigate whether this solution could be integrated in exis-

ting computational models and whether this could help better capture human behavior. Besides, future robotic experiments with these models could tell us more about which mechanisms can be most appropriate for the planning process within the MB system to produce robust and efficient decision-making and behavioral adaptation.

5.1.6 Other currently supervised robotics work

We are currently investigating further such robotic implementations in three different directions. First, we are further testing the applicability of these Computational Neuroscience-based models and principles to Robotics by testing them in more difficult navigation tasks, involving visually-rich and larger environments, and more recent robotic devices such as the PR2 robot designed by Willow Garage (Fig. 5.1). The aim is to determine whether these computational models produce more efficient and adaptive navigation abilities than other engineering approaches to robot navigation. This work is based on the Master internships of Erwan Renaudo, Omar Islas-Ramirez and Scarlett Fres, in the frame of the Emergence(s) Ville de Paris HABOT project coordinated by Benoît Girard, and involving collaborations with Raja Chatila at ISIR, Philippe Gaussier, Arnaud Blanchard and Pierre Delarboulas at the University of Cergy-Pontoise.

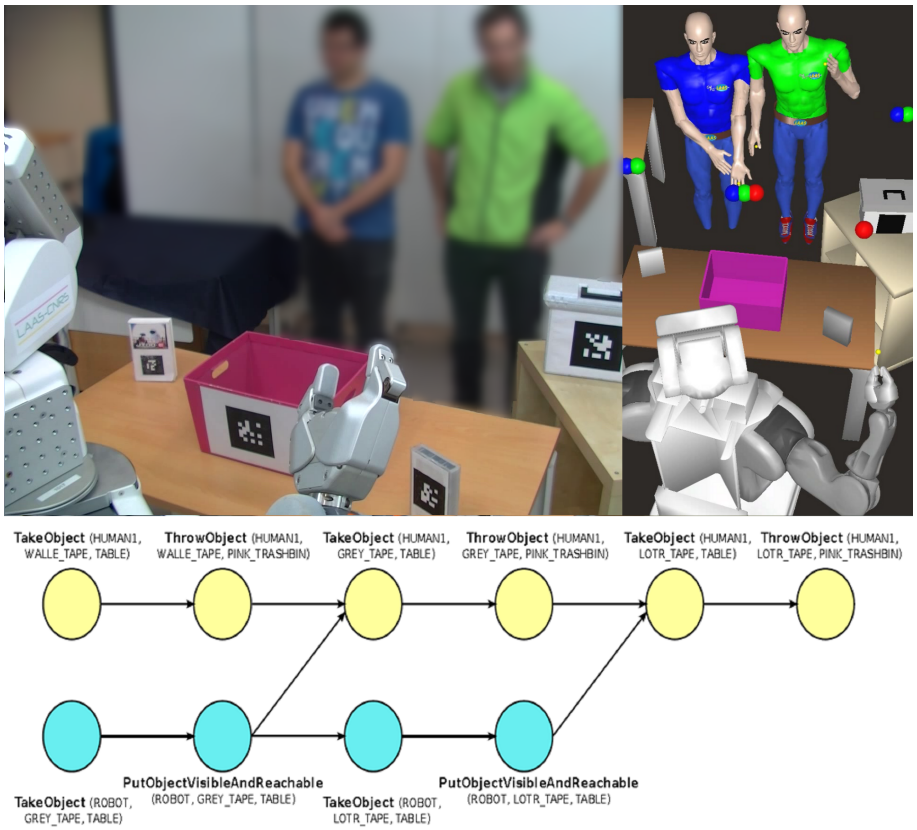


FIGURE 5.2 – *Illustration of the experimental setup with the PR2 Robot at the LAAS-CNRS involving shared action plans during human-robot interaction which will be used to test dual-RL systems models (Based on the work in Alami et al. (2006; 2013), Lemaignan et al. (2012), with permissions)*

Besides, we are also testing the applicability of these models to scena-

rii of human-robot interaction were a shared action plan is required for the coordination of the agents (Alami et al. (2006; 2013), Lemaignan et al. (2012) ; Fig. 5.2). Our learning model could be useful to both (i) autonomously learn a model of the world and deduce efficient action plans, (ii) and enable the robot to acquire habits so as to avoid long and costly computations of the planning process when repetitive actions are performed, for instance in the case of a daily cleaning table task. This work is pursued by PhD student Erwan Renaudo. It involves the collaboration with Raja Chatila and Benoît Girard at ISIR, Rachid Alami and Aurélie Clodic at LAAS-CNRS in Toulouse. Financial support is provided within the ANR *ROBOERGOSUM* project under reference ANR-12-CORD-0030.

Finally, the work of PhD student Nassim Aklil aims at improving the mechanisms used in the model for the coordination of learning systems with recent online budgeted learning techniques from the Machine Learning literature. These techniques provide us with more formal and efficient ways to take into account the computation cost of each learning system, with the aim of improving robotic behavioral performance and assessing the ability of such improved computational models to better reproduce animal behavioral data, making the assumption that the brain attempts to reduce the energy cost when deciding which learning processes to invoke. This work involves the collaboration with Benoît Girard at ISIR, Ludovic Denoyer and Patrick Gallinari at LIP6, and takes place within the framework of the *SMART* LABEX financially supported by French State funds managed by the ANR within the Investissements d'Avenir programme under reference ANR-11-IDEX-0004-02.

5.2 Perspectives and Research Project

The work presented in this manuscript has been subdivided into three main fields of research to which I have contributed and plan to continue contributing throughout the forthcoming years :

1. Computational principles for the coordination of parallel learning processes in animals.
2. Use of the proposed computational models to help better understand experimentally recorded behavioral and neurophysiological data.
3. Implementations of the proposed computational models in behaving robots to help increase their decisional autonomy and learning capacities.

Nevertheless, as highlighted throughout this manuscript, these contributions are not isolated from each other. They talk to each other, and the exploration of possible grounds for their interaction constitutes part the research project sketched in this section. Below are proposed the exploration of four scientific issues related to the coordination of parallel learning processes. The integrative approach adopted to tackle these issues has the potential of contributing to progress in knowledge in both Robotics and Neuroscience. These scientific issues are :

1. On which organization principles should a cognitive architecture be built to enable proper coordination of different learning systems ?

2. How can we model competitive and cooperative interactions between learning systems during multi-strategy navigation ?
3. Which neural signals underly information processing within these different systems during learning under uncertainty ?
4. How should reinforcement learning and motor learning processes interact to enable continuous action / movement acquisition and coordination in the real-world ?

5.2.1 Which cognitive architecture for the coordination of different learning systems ?

Reaching a particular goal in an unprepared environment, or simply surviving while ensuring a sufficient access to required resources, both necessitate the management of several subgoals, potentially antagonistic. They also require to be able to manage events and actions with different priorities and different temporal frequencies. In the studies presented so far, we simplified these issues by assuming a single goal under single fixed motivation so as to explain behavioral phenomena in a single task or in a small set of different situations. In contrast, beyond the learning of particular individual capacities, one has to understand how an agent should react at a given moment to cope with instantaneous constraints while ensuring the achievement of a pursued long-term goal. This long-term goal can either be the survival of an animal or a human, or the mission imposed to a robot by its user.

Several different learning mechanisms can be recruited (latent learning, associative learning, reward-based learning, supervised learning). Different levels of abstractions can be present (perception, analysis of these perceptions, elementary actions, decision-making mechanisms). Hence the need to define a cognitive architecture based on the coordination of different sub-systems to reach a fixed goal.

In Neuroscience researches, having a computational model – and sometimes a hierarchical model – for the description of the interaction between different learning mechanisms can help better capture some experimental data. Of course, the increase in the number of parameters recruited for this high-level model should be penalized to avoid bias in the statistical evaluation of the model's explanatory power (Daw 2011). This approach can also give birth to new hypotheses concerning the interaction between brain regions during behavioral adaptation. This could for instance help us not only say that communication between the prefrontal cortex and the hippocampus has increased at the decision-point during learning (Benchenane et al. 2010), but also which precise changes in information processing may underlie such a dynamics. This approach could also help understand which different computational mechanisms underlie the contribution to learning processes of different anatomical loops linking the prefrontal cortex to the basal ganglia via the thalamus (Alexander et al. 1990, Haber et al. 2000, Keramati and Gutkin 2013). It could also make a link with Computational Neuroscience models integrating homeostatic regulation with reinforcement learning processes (Coninx et al. 2008, Keramati and Gutkin 2011).

In Robotics researches, as mentioned in the introduction, robot control architectures inspired by cognitive architectures proposed in Psychology such as *SOAR* or *ACT-R* (Rosenbloom et al. 1993, Anderson et al. 2004) have long been the subject of intense developments and applications (Alami et al. 1998, Volpe et al. 2001). Such architectures are of particular interest for this project since they include solutions for the management of goals and subgoals, and for the coordination of the planning system with low-level reactive routines. However, these architectures still lack efficient learning abilities and can thus not produce efficient behavioral adaptation in non-stationary environments. Thus investigating how to integrate parallel reinforcement learning mechanisms and their coordination to these architectures could constitute interesting and fruitful lines of research.

This part of my research project will be addressed within the frame of the ANR *ROBOERGOSUM* project under reference ANR-12-CORD-0030, in collaboration with Rachid Alami, Raja Chatila, Aurélie Clodic, Benoît Girard, and Erwan Renaudo.

From the global scientific issue of this project, a set of particular questions of particular interest will be declined :

- How do decision-making and action processes interact ?
- How do instrumental conditioning and Pavlovian conditiong processes interact ?
- How can one learn new elementary capacities and then properly use them ?
- How can one extract abstraction of the elementary actions observed during behavior ? Which level of abstraction ? Which cutting ? Based on which criteria ?
- Should chunked sequences of elementary actions then be considered as habits ? As elements manipulable by the model-based system ? Or both ?
- How can one coordinate goal-oriented learning and latent learning processes ?
- How should changes in motivation and in sub-goals affect the coordination of learning processes ?
- Do changes in motivation or in sub-goals affect neural reinforcement signals and associated neuromodulatory processes ?
- Which implication for anatomical loops linking the prefrontal cortex and the basal ganglia ? How should we model them ?

5.2.2 Cooperation/competition between learning systems for navigation

The objective of this part of my research project is to model the dynamical processes underlying the selection and combination of navigation strategies in the case of complex and non-stationary tasks.

Goal-oriented navigation is a fundamental function in daily life of many species, and probably soon of robots. It implies the ability to acquire knowledge about the environment – such as spatio-temporal dependencies between environmental features – and to use it to apply the most adapted locomotor strategy in a given context (Khamassi and Humphries 2012). Modelling the processes which underlie the dynamical adaptation

of navigation strategies enables to address the question about the nature of processes at stake during the resolution of complex problems. How can animals and robots choose the most adequate strategy or combination of strategies when there are multiple possibilities ?

As we have argued in the Introduction and in Chapter 2, mammals' navigation skills include the ability to alternate between different navigation strategies (e.g. allocentric versus egocentric, cue-guided versus place-based, etc...). At the learning and action selection level, these strategies can relevantly be categorized into model-based and model-free reinforcement learning processes (Khamassi 2007). Such a computational framework, and computational models based on this dichotomy (Dollé et al. 2010; submitted), enable to explain a large body of experimental data, as well as the contribution of different parts of the basal ganglia to navigation (Khamassi and Humphries 2012). Computational models based on these principles and implemented on robotic platforms can capture the richness of the dynamics of interaction with the environment (Caluwaerts et al. 2012b;a).

However, this approach cannot yet account for all navigation strategies observed in animals. For instance route learning, where the model of the world could be independent from the notion of place ; pure motor (praxic) sequence learning, which could be captured by a route over-learning process ; finally metric navigation cannot be accounted for by this approach while it is supposedly frequent in animals and dominating in the field of Engineering of mobile robots. Thus, extending our computational framework to account for these other navigation strategies will constitute a first part of our project.

Besides, mechanisms for the coordination of multiple navigation strategies have been the subject of little modelling work and were often tested in simple tasks requiring a repertoire limited to two strategies. Among these models, some use a mechanism for strategy output fusion – *i.e.* they merge or sum the probability distribution over actions produced by different strategies – (Guazzelli et al. 1998, Girard et al. 2005). Other models use a selection mechanism so that a single strategy controls behavior at each moment (Foster et al. 2000, Chavarriaga et al. 2005). The latter cases require the choice of a strategy selection criterion which should be as much as possible independent from the specificities of the implemented strategies. The computational model developed at ISIR (Dollé et al. 2008; 2010; submitted) and which we used in the robotic implementations presented in Chapter 4 proposed a new selection criterion able to coordinate in an adaptive manner strategies of different types – through the use of a common currency for the evaluation of strategies' performance. This computational model has previously provided us with a computational explanation for the evolution through time of rodent behavior performing tasks limited to two strategies (Pearce et al. 1998, Devan and White 1999). Nevertheless, the model needs to be generalized to tasks involving multiple strategies. Moreover, the hypotheses on possible neural substrates of the different components of the model remain to be tested. Finally, the robotic implementations that we performed with this model showed that it needs to be extended to enable a contextual flexibility of strategy selection (Caluwaerts et al. 2012b).

This part of the project will be addressed in collaboration with Benoît Girard at ISIR for the modelling part and will involve the collaboration with experimentalists : Laure Rondi-Reig (CNRS – UPMC) has conceived a new navigation setup called the *star-maze* to specifically identify the strategy performed by the animal or the combination of strategies which most likely explains its behavior (Rondi-Reig et al. 2006) ; Sidney I. Wiener (CNRS – Collège de France) has developed experimental protocols for the investigation of neural processes underlying the rapid switch between navigation strategies (Battaglia et al. 2008, Peyrache et al. 2009, Catanese et al. 2012). The direct interaction between computational models and experimental data permitted by these collaborations will enable to address a set of specific questions :

- How can we model the sequential route strategy observed in animals in the star-maze ?
- Which principles underlie the transient cooperation or competition between navigation strategies ?
- How do these interactions evolve through time and can the model-based / model-free reinforcement learning computational framework account for this evolution ?
- Which mechanism for strategy selection best explains cases of strategy shifts observed experimentally ?
- Is there a specific substrate underlying the strategy selection mechanism or is this subserved through interactions between neural substrates of each individual strategy ?
- Can manipulations of particular brain areas affect specific learning or strategy selection processes ?
- Can the resulting computational model enable robots to efficiently navigate in a set of very different, unprepared, non-stationary and large environments ?

5.2.3 Neural signals underlying learning under uncertainty

The work presented throughout the manuscript addressed at multiple times the question about the possible neural substrates and their respective mechanisms underlying the coordination of parallel learning processes for behavioral adaptation. Of particular interest were the role of different parts of the striatum – belonging to different anatomical loops between the prefrontal cortex and the basal ganglia – in these learning processes (Khamassi and Humphries 2012), the role of the prefrontal cortex network in cognitive control processes enabling the coordination of learning systems and the regulation of learning parameters (Peyrache et al. 2009, Benchenane et al. 2010, Khamassi et al. 2011b; 2013; 2014), and the role of dopamine signals in the learning process (Bellot et al. 2012; in preparation). Although these contributions tell us more about neural mechanisms of behavioral adaptation, the synergy between these different regions, the differential roles of dopamine on each node of the system, and the fundamental principles that govern these mechanisms are still unknown. Moreover, most of this work concerns experimental situations where the uncertainty about the environment is limited.

Solving problems and adapting to new situations requires coping with

uncertainty about the possible values and consequences of our decisions. In order to reduce uncertainties and optimize decision making, a subject must learn by experience and estimate statistical properties of the environment and actions, such as probabilities of obtaining a valuable outcome or of reaching a more promising environment. During this learning process, uncertainty may arise from two sources that need to be identified : whereas the intrinsic variability of physical or biological phenomena (risk) may be irreducible, uncertainty associated with a reduced knowledge about specific outcomes or processes (ambiguity) may be controlled when more information is collected. Thus, deciding under ambiguity is often opposed to deciding under risk, depending on the information one possesses about the probabilities of possible outcomes. In ecological situations, animals and humans are able to adapt their learning performances according to the changing statistical properties of the environment (Rushworth and Behrens 2008).

Understanding how the brain resolves uncertainty during choice behavior is a fundamental issue in theories of economic decision making (Neuroeconomics), for instance to account for individual differences in attitude towards risk and ambiguity, but also in all fields dealing with adaptive systems (Schultz 2008). Developmental robotics, for instance, are concerned with how cognitive systems can explore and master uncertainty in order to develop (Lungarella et al. 2004). Increased understanding of these processes in animals can inspire the design and control of architectures for artificial agents.

Phasic dopaminergic (DA) activity has been hypothesized to represent a reward prediction error signal and is construed to support learning stimuli and action values in the striatum (Schultz et al. 1997, Bayer and Glimcher 2005, Morris et al. 2006, Roesch et al. 2007). Part of dopamine signalling has been hypothesized to incorporate information about the uncertainty of the environment (Fiorillo and Tobler 2003). However, the relation between such an uncertainty and reinforcement learning processes is a matter of debate (Niv et al. 2005). Moreover, different types of signals have been recently identified in dopamine activity (Matsumoto and Hikosaka 2009, Fiorillo 2013) and dopaminergic cells are functionally dissociable according to their targets (Lammel et al. 2008). This suggests that dopamine projections to prefrontal cortical areas play a different role (Doya 2008), for instance by controlling the construction of task-relevant state representations which would affect the use of model-based decision making and which conversely may differentially affect distinct DA pathways (Takahashi et al. 2011). Finally, it is not clear whether dopamine signals should only impact learning or also influence the action selection process, and different computational propositions have been made to dissociate the effect of different dopamine signals – namely phasic and tonic signals – on these two processes (McClure et al. 2003, Humphries et al. 2012).

In collaboration with several experimental groups, this part of the project thus aims at contributing to the addressment of a set of scientific questions :

– What is the impact of different levels and types of uncertainty on dopamine signalling ?

- Do different dopamine signals subserve different functions in their target structure such as the striatum and prefrontal cortex ?
- How do changes in dopamine signals impact cortical state representations in the model-based system and coordination between the model-free and model-based systems ?
- How does in turn such a coordination impact dopamine signalling ?
- What are the respective roles of phasic and tonic dopamine and how do they modulate parallel learning and decision-making processes in the brain ?

Some of these questions are currently addressed within the frame of a national ANR *Learning Under Uncertainty* project under reference ANR-11-BSV4-006, coordinated by Emmanuel Procyk at INSERM in Lyon, and involving experimental partners in Marseille (Paul Apicella at CNRS) and in Bordeaux (Alain Marchand and Etienne Coutureau at CNRS). Together with these collaboraters, we are using computational models to design experimental protocols that can specifically address some of these issues. In turn, behavioral experiments as well as dopamine electrophysiological recordings should provide us with new data helping to disentangle the computational mechanisms underlying these phenomena.

The issue about the different roles of phasic and tonic dopamine signals are addressed in collaboration with Kevin Gurney at the University of Sheffield and Mark D. Humphries at the University of Manchester. A first computational modelling work has been published, with new experimental predictions raised (Humphries et al. 2012). A collaborative project between the CNRS and the Royal Society has been submitted.

Finally the questions about the mutual influence between dopamine signals and cortical model-based state representations is addressed through a collaboration with Geoffrey Schoenbaum and Matthew R. Roesch from the University of Maryland and the NIDA-NIH, Kenji Doya at Okinawa Institute of Science and Technology, and Alain Marchand and Etienne Coutureau at CNRS in Bordeaux. A collaborative project has been submitted to the Human Frontiers Science Program.

5.2.4 Integration of reinforcement learning and motor control

The work presented in this manuscript mostly focuses on reinforcement learning (RL) and unsupervised learning (UL) processes, and on their corresponding neural substrates in the basal ganglia and prefrontal cortex. The motor control part has not been addressed, which means that the computational models presented are most of the time simplified by manipulating abstract actions without wondering how sequences of muscle activations are learned and organized for the execution of these actions. This part of my research project aims at overcoming this limitation by coordinating RL and UL with supervised learning processes in order to incorporate the motor control part in the computational models. The strong interactions between Robotics and Neuroscience characterizing our approach can reveal particularly fruitful in addressing this issue. Robotic experimentations will indeed play an important part here in that they require the continuous coordination of movements of a physical body in-

teracting with the environment, imposing time and physical constraints on behavior.

As mentioned in the introduction, reinforcement learning algorithms reach their limits when the experimental paradigm takes into account continuous rather than discrete time, state and action information. To overcome these difficulties, several RL algorithms working in the continuous case have been proposed, some dealing with continuous time and space (Doya 2000b), some with continuous state space (Khamassi et al. 2006), and some with continuous action space (Peters and Schaal 2008). The latter study proposed a natural Actor-Critic algorithm – which is model-free – and is based on gradient descent algorithms used as a continuous optimization method for Reinforcement Learning. The links between optimization in the continuous case and RL are the subjects of more and more attention from both Neuroscience and Engineering. This question is central in Engineering work aiming at proposing solutions for the learning of elementary action and motor skills. It is also important in the Neuroscience field concerned with human motor control, action selection or navigation. The addressed processes are for instance fundamental for learning capacities relying on the continuous segmentation of the sensorimotor flow (chunking).

This part of my research project will be strongly based on the local research environment within ISIR, where researches in the AMAC team have contributed to many different aspects of this scientific issue : optimal control theory (Rigoux and Guigon 2012), stochastic multi-objective optimization (Doncieux et al. 2011), robot motor learning (Sigaud et al. 2011), reinforcement learning processes in animals (Humphries et al. 2012) and in robots (Caluwaerts et al. 2012b). This environment will help us better address the following questions :

- How can an agent learn with multiple and continuous perceptions and actions ? What can then guide learning ?
- How are reward and cost integrated in the decision function ?
- How do prefrontal cortex and basal ganglia interact, compete or complete each other for decision-making ?
- Do some fronto-striatal loops also contribute to learning and selection of elementary movements ? Or do these processes remain at an abstract and discrete action level in cortico-basal networks, interacting with other brain structures such as the Cerebellum to subserve the finest grain of movement ?

In collaboration with Benoît Girard and Ignasi Cos at ISIR, we have made a first step in this direction of research by proposing an RL computational model accounting for the influence of the biomechanical cost of movement in the decision-making process observed in humans having to choose between pairs of targets reachable with the hand (Cos et al. 2013). Previous psychophysical experiments performed by Ignasi Cos during his postdoctoral traininig in Paul Cisek's laboratory at the University of Montreal have shown that biomechanical constraints such as the energetic cost of movement were implicitely taken into account in the decisions made by subjects between two possible movements (Cos et al. 2011; 2012). This biomechanical information is integrated with target distance information for optimal decision-making. Subjects would choose to move towards the

closest target, unless the difference in distance is compensated by a higher energy cost in the required movement. Interestingly, most subjects did not report having explicitely included the energetical cost of movements in their decision process.

We have shown that model-based RL mechanisms could account for these data by assuming the integration of biomechanical costs within the model of the world during a learning phase prior to the experiment, during a developmental motor babbling phase (Cos et al. 2013). Interestingly, additional analyses on human behavioral data confirmed that the integration of biomechanical costs in decision-making were present from the very first trials of the experiment. Moreover, the model enabled us to draw a set of experimental predictions about expected human behavior in cases where decision time is limited. This computational model will constitute a basis for future theoretical developments planned for this part of my research project.

Conclusion

The presented research project aims at understanding and formalizing fundamental principles and methods that underlie animals' and robots' behavioral adaptation capabilities. Fundamental questions are raised such as how can animals and robots understand their own actions and their consequences ? How can they learn from their errors in an open-ended fashion ? How can they exhibit curious and exploratory behaviors ? How can they monitore their own performance as well as the environment so as to regulate their learning parameters, and so as to learn to autonomously shift to the decisional mode which is the most appropriate for the current situation and for the current level of uncertainty in order to more efficiently adapt, and acquire new knowledge and new skills ?

Neuroscience data constitute a source of inspiration for Robotics concerning the way the brain's cognitive architecture coordinates different hierarchical levels of decision-making, and how it selects and integrates relevant information for efficient adaptive behavior. It also provides experimental tasks or scenarii that enable to isolate particular behaviors and learning abilities. The work presented in this manuscript describes how computational principles for the coordination of parallel learning processes can account for biological data, and can be implemented on robotic platforms to reproduce animal laboratory tasks. The multidisciplinary nature of the adopted approach makes the questions addressed be potential contributions to artificial systems with the aim of improving robots' autonomy and adaptation capabilities, but also to behavioral and brain research. The results could be in turn of interest to Neuroscience and Cognitive Science because they assess the robustness of Computational Neuroscience models tested in the real world.

References

R. Alami, R. Chatila, S. Fleury, M. Ghallab, and F. Ingrand. An architecture for autonomy. *International Journal of Robotics Research*, 17(4) :315–337, 1998. (Cited on pages 11 and 200.)

R. Alami, A. Clodic, V. Montreuil, E.A. Sisbot, and R. Chatila. Toward human-aware robot task planning. In *AAAI Spring Symposium : To Boldly Go Where No Human-Robot Team Has Gone Before*, pages 39–46. 2006. (Cited on pages vii, 11, 197, and 198.)

R. Alami, M. Warnier, J. Guitton, S. Lemaignan, and E.A. Sisbot. When the robot considers the human... In O. Khatib and H. Christensen, editors, *Proceedings of the 15th International Symposium on Robotics Research*. Springer, 2013. (Cited on pages vii, 197, and 198.)

G.E. Alexander, M.D. Crutcher, and M.R. DeLong. Basal ganglia-thalamocortical circuits : parallel substrates for motor, oculomotor, "prefrontal" and "limbic" functions. *Progress in Brain Research*, 85 :119–146, 1990. (Cited on page 199.)

W.H. Alexander and O. Sporns. An embodied model of learning, plasticity, and reward. *Adaptive Behavior*, 10 :143–159, 2002. (Cited on page 10.)

J.R. Anderson, D. Bothell, M.D. Byrne, S. Douglas, C. Lebiere, and Y. Qin. An integrated theory of the mind. *Psychological Review*, 111(4) :1036–1060, 2004. (Cited on page 200.)

A. Angeli, D. Filliat, S. Doncieux, and J.-A. Meyer. A fast and incremental method for loop-closure detection using bags of visual words. *IEEE Transactions On Robotics, Special Issue on Visual SLAM*, 24(5), 2008. (Cited on page 12.)

M. Arbib, G. Metta, and P. van der Smagt. Neurorobotics : From vision to action. In B. Siciliano and O. Khatib, editors, *Handbook of robotics*, pages 1453–1480. Springer-Verlag, Berlin, 2008. (Cited on page 12.)

A. Arleo and W. Gerstner. Spatial cognition and neuro-mimetic navigation : a model of hippocampal place cell activity. *Biological Cybernetics*, 83(3) :287–299, 2000. (Cited on page 12.)

A. Arleo, F. Smeraldi, and W. Gerstner. Cognitive navigation based on nonuniform gabor space sampling, unsupervised growing networks, and reinforcement learning. *IEEE Transactions on Neural Networks*, 15 :639–652, 2004. (Cited on page 10.)

F.G. Ashby, B.O. Turner, and J.C. Horvitz. Cortical and basal ganglia contributions to habit learning and automaticity. *Trends in Cognitive Science*, 14 :208–215, 2010. (Cited on page 6.)

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47 :235–256, 2002. (Cited on pages 9 and 194.)

D. Badre and M. D'Esposito. Is the rostro-caudal axis of the frontal lobe hierarchical ? *Nature Reviews Neuroscience*, 10 :659–669, 2009. (Cited on page 6.)

G. Baldassarre. A modular neural-network model of the basal ganglia's role in learning and selecting motor behaviours. *Cognitive Systems Research*, 3(1) :5–13, 2002. (Cited on page 5.)

B.W. Balleine and J.P. O'Doherty. Human and rodent homologies in action control : corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, 35(1) :48–69, 2010. (Cited on pages 5, 6, and 192.)

A. Barrera, A. Caceres, A. Weitzenfeld, and V. Ramirez-Amaya. Comparative experimental studies on spatial memory and learning in rats and robots. *Journal of Intelligent and Robotic Systems*, 63(3–4) :361–397, 2011. (Cited on page 13.)

F.P. Battaglia, A. Peyrache, M. Khamassi, and S.I. Wiener. Spatial decisions and neuronal activity in hippocampal projection zones in prefrontal cortex and striatum. In S. Mizumori, editor, *Hippocampal Place Fields : Relevance to Learning and Memory*, chapter 18, pages 289–311. Oxford University Press, 2008. (Cited on pages 9 and 202.)

H.M. Bayer and P.W. Glimcher. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, 47(1) :129–141, 2005. (Cited on pages 5 and 203.)

T.E. Behrens, M.W. Woolrich, M.E. Walton, and M.F. Rushworth. Learning the value of information in an uncertain world. *Nature Neuroscience*, 10 (9) :1214–1221, 2007. (Cited on page 10.)

J. Bellot, O. Sigaud, and M. Khamassi. Neuro-inspired navigation strategies shifting for robots : Integration of a multiple landmark taxon strategy. In T. Ziemke, C. Balkenius, and J. Hallam, editors, *From Animals to Animats : Proceedings of the 12th International Conference on Adaptive Behaviour (SAB 2012)*, pages 289–298. Springer, 2012. (Cited on pages 18 and 202.)

J. Bellot, O. Sigaud, M.R. Roesch, G. Schoenbaum, B. Girard, and M. Khamassi. What do vta dopamine neurons encode : value, rpe or other behavior ? in preparation. (Cited on pages 6, 17, 69, 194, and 202.)

K. Benchenane, A. Peyrache, M. Khamassi, P.L. Tierney, Y. Gioanni, F.P. Battaglia, and S.I. Wiener. Coherent theta oscillations and reorganization of spike timing in the hippocampal- prefrontal network upon learning. *Neuron*, 66(6) :921–936, 2010. (Cited on pages 9, 147, 199, and 202.)

P. Biber and T. Duckett. Dynamic maps for long-term operation of mobile service robots. In *Proceedings of the ECML-98 Workshop on Upgrading Learning to Meta-Level : Model Selection and Data Transformatio*, pages 11–17. 2005. (Cited on page 12.)

P. Brazdil. Data transformation and model selection by experimentation and meta-learning. In *Robotics : science and systems*, pages 17–24. 1998. (Cited on page 9.)

R. Brooks. A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, RA-2 :14–23, 1986. (Cited on page 11.)

A. Brovelli, N. Laksiri, B. Nazarian, M. Meunier, and Boussaoud D. Understanding the neural computations of arbitrary visuomotor learning through fmri and associative learning theory. *Cerebral Cortex*, 18(1) : 1485–1495, 2008. (Cited on page 5.)

A. Brovelli, B. Nazarian, M. Meunier, and D. Boussaoud. Differential roles of caudate nucleus and putamen during instrumental learning. *NeuroImage*, 57(4) :1580–90, 2011. (Cited on page 192.)

K. Caluwaerts, A. Favre-Félix, M. Staffa, S. N'Guyen, C. Grand, B. Girard, and M. Khamassi. Neuro-inspired navigation strategies shifting for robots : Integration of a multiple landmark taxon strategy. In T.J. Prescott, N.F. Lepora, A. Mura, and P.F.M.J. Verschure, editors, *1st Living Machines Conference, Lecture Notes in Artificial Intelligence 7375*, pages 62–73. Springer, 2012a. (Cited on pages 13, 18, and 201.)

K. Caluwaerts, M. Staffa, N'Guyen. S., C. Grand, L. Dollé, A. Favre-Félix, B. Girard, and M. Khamassi. A biologically inspired meta-control navigation system for the psikharpax rat robot. *Bioinspiration and Biomimetics*, 7(2) :025009, 2012b. (Cited on pages 8, 13, 18, 147, 195, 201, and 205.)

J. Catanese, E. Cerasti, M. Zugaro, A. Viggiano, and S.I. Wiener. Dynamics of decision-related activity in hippocampus. *Hippocampus*, 22(9) :1901–1911, 2012. (Cited on page 202.)

R. Chatila, R. Alami, B. Degallaix, and H. Laruelle. Integrated planning and execution control of autonomous robot actions. In *Proceedings of the IEEE International Conference on Robotics and Automation, ICRA'92*, pages 2689–2696. 1992. (Cited on page 11.)

R. Chavarriaga, T. Strösslin, D. Sheynikhovich, and W. Gerstner. A computational model of parallel navigation systems in rodents. *Neuroinformatics*, 3 :223–241, 2005. (Cited on page 201.)

A.G.E. Collins and M.J. Frank. How much of reinforcement learning is working memory, not reinforcement learning ? a behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, 35 : 1024–1035, 2012. (Cited on page 192.)

A.G.E. Collins and E. Koechlin. Reasoning, learning, and creativity : frontal lobe function and human decision-making. *PLoS Biology*, 10(3) : e1001293, 2012. (Cited on page 5.)

A. Coninx, A. Guillot, and B. Girard. Adaptive motivation in a biomimetic action selection mechanism. In *Proceedings of the NeuroComp Conference 2008*, pages 158–162, Marseille, France, 2008. (Cited on page 199.)

L.H. Corbit and B.W. Balleine. The general and outcome-specific forms of pavlovian-instrumental transfer are differentially mediated by the nucleus accumbens core and shell. *The Journal of Neuroscience*, 31(33) : 11786–11794, 2011. (Cited on page 192.)

G. Corrado and K. Doya. Understanding neural coding through the model-based analysis of decision making. *Journal of Neuroscience*, 27 (31) :8178–8180, 2007. (Cited on page 5.)

I. Cos, N. Bélanger, and P. Cisek. The influence of predicted arm biomechanics on decision making. *Journal of Neurophysiology*, 105 :3022–3033, 2011. (Cited on page 205.)

I. Cos, F. Medleg, and P. Cisek. The modulatory influence of end-point controllability on decisions between actions. *Journal of Neurophysiology*, 108 :1764–1780, 2012. (Cited on page 205.)

I. Cos, M. Khamassi, and B Girard. Modeling the learning of biomechanics and visual planning for decision-making of motor actions. *Journal of Physiology*, 107(5) :399–408, 2013. (Cited on pages 205 and 206.)

N.D. Daw. Affect, learning and decision making, attention and performance. vol. xxiii. chapter Trial-by-trial data analysis using computational models, pages 3–38. Oxford University Press, Oxford, UK, 2011. (Cited on pages 5 and 199.)

N.D. Daw, Y. Niv, and P. Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12) :1704–1711, 2005. (Cited on pages 6, 7, 192, 195, and 196.)

N.D. Daw, J.P. O'Doherty, P. Dayan, B. Seymour, and R.J. Dolan. Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095) :876–879, 2006. (Cited on page 5.)

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

P. Dayan and B.W. Balleine. Reward, motivation, and reinforcement learning. *Neuron*, 36(2) :285–298, 2002. (Cited on page 6.)

P. Dayan and Y. Niv. Reinforcement learning and the brain : The good, the bad and the ugly. *Current Opinion in Neurobiology*, 18(2) :185–196, 2008. (Cited on page 3.)

B.D. Devan and N.M. White. Parallel information processing in the dorsal striatum : relation to hippocampal function. *The Journal of Neuroscience*, 19(7) :2789–2798, 1999. (Cited on page 201.)

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

A. Dickinson. Actions and habits : The development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 308(1135) :67–78, 1985. (Cited on page 6.)

L. Dollé, M. Khamassi, B. Girard, A. Guillot, and R. Chavarriaga. Analyzing interactions between navigation strategies using a computational model of action selection. In *Spatial Cognition Conference, Lecture Notes in Computer Science 5248*, pages 71–86. Springer, 2008. (Cited on pages 8, 195, and 201.)

L. Dollé, D. Sheynikhovich, B. Girard, R. Chavarriaga, and A. Guillot. Path planning versus cue responding : a bioinspired model of switching between navigation strategies. *Biological Cybernetics*, 103(4) :299–317, 2010. (Cited on pages 8, 195, 196, and 201.)

L. Dollé, R. Chavarriaga, M. Khamassi, and A. Guillot. Interactions between spatial strategies producing generalization gradient and blocking : a computational approach. *Psychological Reviews*, submitted. (Cited on pages 8, 195, and 201.)

S. Doncieux, N. Bredeche, and J-B. Mouret, editors. *New Horizons in Evolutionary Robotics*. Springer-Verlag, Berlin Heidelberg, 2011. (Cited on pages 9 and 205.)

K. Doya. Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology*, 10 :732–739, 2000a. (Cited on pages vii, 3, and 5.)

K. Doya. Reinforcement learning in continuous time and space. *Neural Computation*, 12(1) :219–245, 2000b. (Cited on page 205.)

K. Doya. Metalearning and neuromodulation. *Neural Networks*, 15(4-6) : 495–506, 2002. (Cited on page 8.)

K. Doya. Modulators of decision making. *Nature Neuroscience*, 11(4) :410–416, 2008. (Cited on pages 194 and 203.)

A. Faure, U. Haberland, F. Condé, and N. El Massioui. Lesion to the nigrostriatal dopamine system disrupts stimulus-response habit formation. *Journal of Neuroscience*, 25 :2771–2780, 2005. (Cited on page 5.)

C.D. Fiorillo. Two dimensions of value : dopamine neurons represent reward but not aversiveness. *Science*, 341(6145) :546–549, 2013. (Cited on pages 194 and 203.)

C.D. Fiorillo and W. Tobler, P.N. Schultz. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, 299(5614) :1898–902, 2003. (Cited on page 203.)

J. Fix, N. Rougier, and F. Alexandre. From physiological principles to computational models of the cortex. *Journal of Physiology Paris*, 101(1–3) :32–39, 2007. (Cited on page 3.)

J.G. Fleischer, J.A. Gally, G.M. Edelman, and J.L. Krichmar. Retrospective and prospective responses arising in a modeled hippocampus during maze navigation by a brain-based device. *Proceedings of the National Academy of Science*, 104(9) :3556–3561, 2007. (Cited on page 13.)

J. Folkesson and H.I. Christensen. Integrated planning and execution control of autonomous robot actions. In *Proceedings of the IEEE International Conference on Robotics and Automation, ICRA'04*, pages 383–390. 2004. (Cited on page 12.)

D. Foster, R. Morris, and P. Dayan. Models of hippocampally dependent navigation using the temporal difference learning rule. *Hippocampus*, 10 :1–16, 2000. (Cited on page 201.)

M.J. Frank. Dynamic dopamine modulation in the basal ganglia : a neurocomputational account of cognitive deficits in medicated and nonmedicated parkinsonism. *Journal of Cognitive Neuroscience*, 17(1) :51–72, 2005. (Cited on pages 5 and 193.)

M.J. Frank, B.B. Doll, J. Oas-Terpstra, and F. Moreno. Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience*, 12(8) :1062–1068, 2009. (Cited on page 5.)

H. Frezza-Buet, N. Rougier, and F. Alexandre. Integration of biologically inspired temporal mechanisms into a cortical framework for sequence processing. In R. Sun and C.L. Giles, editors, *Sequence Learning : Paradigms, Algorithms, and Applications, LNAI 1828*, pages 321–348. Springer, 2001. (Cited on page 12.)

C. Giovannangeli and P. Gaussier. Autonomous vision-based navigation : Goal-oriented action planning by transient states prediction, cognitive map building, and sensory-motor learning. In *Proceedings of the International Conference on Intelligent Robots and Systems*, volume 1, pages 281–297. University of California Press, 2008. (Cited on page 13.)

B. Girard, D. Filliat, J.A. Meyer, A. Berthoz, and A. Guillot. Integration of navigation and action selection functionalities in a computational model of corticobasal gangliathalamocortical loops. *Adaptive Behavior*, 13(2) : 115–130, 2005. (Cited on page 201.)

C. Giraud-Carrier, P. Brazdil, and R. Vilalta. Introduction to the special issue on meta-learning. *Machine Learning*, 54(3) :187–193, 2004. (Cited on page 8.)

J. Gläscher, N.D. Daw, P. Dayan, and J.P. O'Doherty. States versus rewards : dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66 :585–595, 2010. (Cited on page 6.)

A.M. Graybiel. Habits, rituals, and the evaluative brain. *Anuual Review of Neuroscience*, 31 :359–387, 2008. (Cited on page 6.)

A. Guazzelli, F.J. Corbacho, M. Bota, and M.A. Arbib. Affordances, motivations and the world graph theory. *Adaptive Behavior : Special issue on biologically inspired models of spatial navigation*, 6(34) :435–471, 1998. (Cited on page 201.)

K. Gurney, T.J. Prescott, and P. Redgrave. A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biological Cybernetic*, 84(6) :401–410, 2001. (Cited on page 5.)

S.N. Haber, J.L. Fudge, and N.R. McFarland. Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *The Journal of Neuroscience*, 20(6) :2369–82, 2000. (Cited on pages 194 and 199.)

M.E. Hasselmo. A model of prefrontal cortical mechanisms for goal-directed behavior. *Journal of Cognitive Neuroscience*, 17(7) :1115–1129, 2005. (Cited on page 3.)

J.C. Houk, J.L. Adams, and A.G. Barto. A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk, J. L. Davis, and D. G. Beiser, editors, *Models of Information Processing in the Basal Ganglia*, pages 249–271. The MIT Press, Cambridge, MA, 1995. (Cited on page 5.)

M.W. Howe, P.L. Tierney, S.G. Sandberg, P.E. Phillips, and A.M. Graybiel. Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. *Nature*, 500(7464) :575–579, 2013. (Cited on page 194.)

M.D. Humphries and T.J. Prescott. The ventral basal ganglia, a selection mechanism at the crossroads of space, strategy, and reward. *Progress in Neurobiology*, 90 :385–417, 2010. (Cited on page 5.)

M.D. Humphries, M. Khamassi, and K. Gurney. Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. *Frontiers in Neuroscience*, 6 :9, 2012. (Cited on pages 5, 195, 203, 204, and 205.)

Q.J. Huys, N. Eshel, E. O'Nions, L. Sheridan, P. Dayan, and J.P. Roiser. Bonsai trees in your head : how the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Computational Biology*, 8(3) : e1002410, 2012. (Cited on page 196.)

S. Ishii, W. Yoshida, and J. Yoshimoto. Control of exploitation-exploration meta-parameter in reinforcement learning. *Neural Networks*, 15(4-6) : 665–687, 2002. (Cited on page 9.)

M. Ito and K. Doya. Validation of decision-making models and analysis of decision variables in the rat basal ganglia. *Journal of Neuroscience*, 29 (31) :9861–9874, 2009. (Cited on page 5.)

Makoto Ito and Kenji Doya. Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit. *Current Opinion in Neurobiology*, 21 :368–373, 2011. (Cited on page 6.)

D. Joel, Y. Niv, and E. Ruppin. Actor-critic models of the basal ganglia : new anatomical and computational perspectives. *Neural Networks*, 15 (4-6) :535–547, 2002. (Cited on page 5.)

O. Kanoun, J.P. Laumond, and E. Yoshida. Planning foot placements for a humanoid robot : A problem of inverse kinematics. *International Journal of Robotics Research*, 30(4) :476–485, 2011. (Cited on page 11.)

M. Kawato, S. Kuroda, and N. Schweighofer. Cerebellar supervised learning revisited : biophysical modeling and degrees-of-freedom control. *Current Opinion in Neurobiology*, 21(5) :791–800, 2011. (Cited on page 2.)

M. Keramati and B. Gutkin. Imbalanced decision hierarchy in addicts emerging from drug-hijacked dopamine spiraling circuit. *PLoS ONE*, 8 (4) :e61489, 2013. (Cited on pages 3 and 199.)

M. Keramati and B.S. Gutkin. A reinforcement learning theory for homeostatic regulation. In J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, editors, *NIPS*, pages 82–90, 2011. (Cited on page 199.)

M. Keramati, A Dezfouli, and P. Piray. Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Computational Biology*, 7(5) :e1002055, 2011. (Cited on pages 7, 195, and 196.)

M. Khamassi. *Complementary roles of the rat prefrontal cortex and striatum in reward-based learning and shifting navigation strategies*. PhD thesis, Université Pierre et Marie Curie, 2007. (Cited on pages 5, 15, and 201.)

M. Khamassi and M.D. Humphries. Integrating cortico-limbic-basal ganglia architectures for learning model-based and model-free navigation strategies. *Frontiers in Behavioral Neuroscience*, 6 :79, 2012. (Cited on pages 4, 8, 17, 19, 191, 200, 201, and 202.)

M. Khamassi, L. Lacheze, B. Girard, A. Berthoz, and A. Guillot. Actor-critic models of reinforcement learning in the basal ganglia : from natural to arificial rats. *Adaptive Behavior*, 13 :131–148, 2005. (Cited on pages 5 and 10.)

M. Khamassi, L.-E. Martinet, and A. Guillot. Combining self-organizing maps with mixtures of experts : application to an actor-critic model of reinforcement learning in the basal ganglia. In *From Animals to Animats 9, LNAI 4095*, pages 394–405. Berlin, Heidelberg : Springer-Verlag, 2006. (Cited on pages 5, 10, and 205.)

M. Khamassi, A.B. Mulder, E. Tabuchi, V. Douchamps, and S.I. Wiener. Anticipatory reward signals in ventral striatal neurons of behaving rats. *European Journal of Neuroscience*, 28 :1849–1866, 2008. (Cited on page 5.)

M. Khamassi, S. Lallée, P. Enel, E. Procyk, and P.F. Dominey. Robot cognitive control with a neurophysiologically inspired reinforcement learning model. *Frontiers in Neurorobotics*, 5 :1, 2011a. (Cited on page 10.)

M. Khamassi, C. Wilson, R. Rothé, R. Quilodran, P.F. Dominey, and E. Procyk. Meta-learning, cognitive control, and physiological interactions between medial and lateral prefrontal cortex. In R.B. Mars, J. Sallet, M.F.S. Rushworth, and N. Yeung, editors, *Neural Basis of Motivational and Cognitive Control*, pages 351–370. Cambridge, MA : MIT Press, 2011b. (Cited on pages 9 and 202.)

M. Khamassi, P. Enel, P.F. Dominey, and E. Procyk. Medial prefrontal cortex and the adaptive regulation of reinforcement learning parameters. *Progress in Brain Research*, 202 :441–464, 2013. (Cited on pages 9 and 202.)

M. Khamassi, R. Quilodran, P. Enel, P.F. Dominey, and E. Procyk. Behavioral regulation and the modulation of information coding in the lateral prefrontal and cingulate cortex. *Cerebral Cortex*, 2014. in press. (Cited on pages 5, 10, 17, 69, 193, and 202.)

S. Killcross and E. Coutureau. Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral Cortex*, 13 :400–408, 2003. (Cited on page 7.)

J. Kober and J. Peters. Policy search for motor primitives in robotics. *Machine Learning*, 84 :171–203, 2011. (Cited on page 11.)

E. Koechlin and C. Summerfield. An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences*, 11(6) :22935, 2007. (Cited on page 9.)

E. Koechlin, C. Ody, and F. Kouneiher. The architecture of cognitive control in the human prefrontal cortex. *Science*, 302(5648) :1181–1185, 2003. (Cited on page 6.)

J.L. Krichmar and G.M. Edelman. Machine psychology : Autonomous behavior, perceptual categorization, and conditioning in a brain-based device. *Cerebral Cortex*, 12 :818–830, 2002. (Cited on page 10.)

J.L. Krichmar, A.K. Seth, D.A. Nitz, J.G. Fleischer, and G.M. Edelman. Spatial navigation and causal analysis in a brain-based device modeling cortical-hippocampal interactions. *Neuroinformatics*, 3(3) :147–169, 2005. (Cited on page 13.)

S. Lammel, A. Hetzel, O. Hackel, I. Jones, B. Liss, and J. Roeper. Unique properties of mesoprefrontal neurons within a dual mesocorticolimbic dopamine system. *Neuron*, 57 :760–773, 2008. (Cited on pages 194 and 203.)

N. Lavesson and P. Davidsson. Quantifying the impact of learning algorithm parameter tuning. *AAAI National Conference on Artificial Intelligence*, 21(1) :395–400, 2006. (Cited on page 8.)

S. Lemaignan, R. Ros, E.A. Sisbot, R. Alami, and M. Beetz. Grounding the interaction : anchoring situated discourse in everyday human-robot interaction. *International Journal of Social Robotics*, 4(2) :181–199, 2012. (Cited on pages vii, 197, and 198.)

N. Lepora, P. Verschure, and T. Prescott. The state of the art in biomimetics. *Bioinspiration and Biomimetics*, 8 :013001, 2013. (Cited on page 13.)

F. Lesaint, O. Sigaud, S.B. Flagel, T.E. Robinson, and M. Khamassi. Modelling individual differences observed in pavlovian autoshaping in rats using a dual learning systems approach and factored representations. *PLoS Computational Biology*, 10(2) :e1003466, 2014. (Cited on pages 8, 17, 19, and 192.)

F. Lesaint, O. Sigaud, J.J. Clark, S.B. Flagel, and M. Khamassi. Experimental predictions drawn from a computational model of sign-trackers and goal-trackers. *Journal of Physiology Paris*, submitted. (Cited on pages 18 and 192.)

J. Liénard and B. Girard. A biologically constrained model of the whole basal ganglia addressing the paradoxes of connections and selection. *Journal of Computational Neuroscience*, pages 1–24, 2013. (Cited on page 193.)

M. Likhachev, M. Kaess, and R.C. Arkin. Learning behavioral parameterization using spatio-temporal case-based reasoning. In *Proceedings of the IEEE International Conference on Robotics and Automation, ICRA'02*, pages 1282–1289. 2002. (Cited on page 11.)

M. Lungarella, G. Metta, R. Pfeifer, and G. Sandini. Developmental robotics : a survey. *Connection Science*, 15(4) :151–190, 2004. (Cited on pages 13 and 203.)

T. Maia and M.J. Frank. From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience*, 14(2) :154–162, 2011. (Cited on page 3.)

A. Marchand, V. Fresno, M. Khamassi, and E. Coutureau. Dopaminergic modulation of the exploration level in a non-stationary probabilistic task. In *FENS Abstract*. 2014. (Cited on page 195.)

L.-E. Martinet, D. Sheynikhovich, K. Benchenane, and A. Arleo. Spatial learning and action planning in a prefrontal cortical network model. *PLoS Computational Biology*, 7(5) :e1002045, 2011. (Cited on page 3.)

M. Matsumoto and O. Hikosaka. Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature*, 459(7248) : 837–841, 2009. (Cited on pages 5, 194, and 203.)

S.M. McClure, N.D. Daw, and P.R. Montague. A computational substrate for incentive salience. *Trends in Neurosciences*, 26(8) :423–428, 2003. (Cited on page 203.)

J.-A. Meyer and A. Guillot. Biologically-inspired robots. In B. Siciliano and O. Khatib, editors, *Handbook of robotics*, pages 1395–1422. Springer-Verlag, Berlin, 2008. (Cited on page 12.)

M. Milford and G. Wyeth. Persistent navigation and mapping using a biologically inspired slam system. *The International Journal of Robotics Research*, 29(9) :1131–1153, 2010. (Cited on page 13.)

E.K. Miller and J.D. Cohen. An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24 :167202, 2001. (Cited on page 9.)

J. Minguez, F. Lamiraux, and J.P. Laumond. Motion planning and obstacle avoidance. In B. Siciliano and O. Khatib, editors, *Handbook of Robotics*, pages 827–852. Springer-Verlag, 2008. (Cited on page 11.)

B. Mitchinson, M.J. Pearson, A.G. Pipe, and T.J. Prescott. Biomimetic robots as scientific models : A view from the whisker tip, 2011. (Cited on page 13.)

M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit. Fast-slam : A factored solution to the simultaneous localization and mapping problem. In *Proceedings of the AAAI National Conference on Artificial Intelligence*, pages 593–598. 2002. (Cited on page 12.)

J. Morimoto and K. Doya. Acquisition of stand-up behavior by a real robot using hierarchical reinforcement learning. *Robotics and Autonomous Systems*, 36 :37–51, 2001. (Cited on page 10.)

G. Morris, A. Nevet, D. Arkadir, E. Vaadia, and H. Bergman. Midbrain dopamine neurons encode decisions for future action. *Nature Neuroscience*, 9(8) :1057–1063, 2006. (Cited on pages 5 and 203.)

P. Moutarlier and R. Chatila. Stochastic multisensory data fusion for mobile robot location and environment modeling. *5th International Symposium on Robotics Research*, pages 85–94, 1985. (Cited on page 12.)

S. N'Guyen, P. Pirim, and J.-A. Meyer. Texture discrimination with artificial whiskers in the robot- rat psikharpax. In *Biomedical Engineering Systems and Technologies : Third International Joint Conference, BIOSTEC 2010*, pages 127–152. Valencia, Spain, 2011. (Cited on page 13.)

Y. Niv, M.O. Duff, and P. Dayan. Dopamine, uncertainty and td learning. *Behavioral and Brain Functions*, 4 :1–6, 2005. (Cited on page 203.)

J. O'Doherty, P. Dayan, J. Schultz, R. Deichmann, K. Friston, and R.J. Dolan. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, 304 :452–454, 2004. (Cited on page 5.)

P.-Y. Oudeyer and F. Kaplan. What is intrinsic motivation ? a typology of computational approaches. *Frontiers in Neurorobotics*, 1 :6, 2007. (Cited on page 13.)

S. Palminteri, M. Lebreton, Y. Worbe, D. Grabli, A. Hartmann, and M. Pessiglione. Pharmacological modulation of subliminal learning in parkinson's and tourette's syndromes. *Proceedings of the National Academy of Science*, 106(45) :19179–19184, 2009. (Cited on page 5.)

A. Parent and L.N. Hazrati. Functional anatomy of the basal ganglia. i. the cortico-basal ganglia-thalamo-cortical loop. *Brain Research Reviews*, 20(1) :91–127, 1995a. (Cited on page 193.)

A. Parent and L.N. Hazrati. unctional anatomy of the basal ganglia. ii. the place of subthalamic nucleus and external pallidum in basal ganglia circuitry. *Brain Research Reviews*, 20(1) :128–154, 1995b. (Cited on page 193.)

J.M. Pearce, A.D. Roberts, and M. Good. Hippocampal lesions disrupt navigation based on cognitive maps but not heading vectors. *Nature*, 396(6706) :75–77, 1998. (Cited on page 201.)

M. Pessiglione, B. Seymour, G. Flandin, R.J. Dolan, and C.D. Frith. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, 442(7106) :1042–1045, 2006. (Cited on page 5.)

J. Peters and S. Schaal. Policy gradient methods for robotics. In *Proceedings of the IEEE International Conference on Intelligent Robotics Systems (IROS)*, pages 2219–2225. 2006. (Cited on page 7.)

J. Peters and S. Schaal. Reinforcement learning of motor skills with policy gradients. *Neural networks*, 21(4) :682–697, 2008. (Cited on pages 11 and 205.)

A. Peyrache, M. Khamassi, K. Benchenane, S.I. Wiener, and F.P. Battaglia. Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. *Nature Neuroscience*, 12(7) :919–926, 2009. (Cited on pages 9, 147, and 202.)

A. Peyrache, K. Benchenane, M. Khamassi, S.I. Wiener, and F.P. Battaglia. Principal component analysis of ensemble recordings reveals cell assemblies at high temporal resolution. *Journal of Computational Neuroscience*, 29(1–2) :309–325, 2010a. (Cited on page 9.)

A. Peyrache, K. Benchenane, M. Khamassi, S.I. Wiener, and F.P. Battaglia. Sequential reinstatement of neocortical activity during slow oscillations depends on cells' global activity. *Frontiers in Systems Neuroscience*, 3 :18, 2010b. (Cited on page 9.)

R. Pfeifer, M. Lungarella, and F. Iida. Self-organization, embodiment, and biologically inspired robotics. *Science*, 318 :1088–1093, 2007. (Cited on page 12.)

P. Redgrave, T.J. Prescott, and K. Gurney. The basal ganglia : a vertebrate solution to the selection problem ? *Neuroscience*, 89(4) :1009–1023, 1999. (Cited on page 5.)

E. Renaudo, B Girard, R. Chatila, and M. Khamassi. Design of a decision architecture for habit learning in robots. In A. Duff, N.F. Lepora, A. Mura, T.J. Prescott, and P.F.M.J. Verschure, editors, *3rd Living Machines Conference, Lecture Notes in Artificial Intelligence 8608*, pages 249–260. Springer, 2014. (Cited on pages 18, 147, and 196.)

R.A. Rescorla and A.R. Wagner. Classical conditioning ii : Current research and theory. chapter A theory of Pavlovian conditioning : variations in the effectiveness of reinforcement and nonreinforcement, pages 64–99. Appleton-Century-Crofts, New-York, 1972. (Cited on page 4.)

J.N. Reynolds, B.I. Hyland, and J.R. Wickens. A cellular mechanism of reward-related learning. *Nature*, 413 :67–70, 2001. (Cited on page 5.)

L. Rigoux and E. Guigon. A model of reward- and effort-based optimal decision making and motor control. *PLOS Computational Biology*, 8(10) : e1002716, 2012. (Cited on page 205.)

M.R. Roesch, D.J. Calu, and G. Schoenbaum. Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nature Neuroscience*, 10(12) :1615–1624, 2007. ISSN 1097–6256. (Cited on pages 5 and 203.)

L. Rondi-Reig, G.H. Petit, C. Tobin, S. Tonegawa, J. Mariani, and A. Berthoz. Impaired sequential egocentric and allocentric memories in forebrain-specific-nmda receptor knock-out mice during a new task dissociating strategies of navigation. *The Journal of Neuroscience*, 26(15) : 4071–4081, 2006. (Cited on page 202.)

P. Rosenbloom, J. Laird, and A. Newell, editors. *The Soar Papers : Research on Integrated Intelligence*. MIT Press, Cambridge, Massachusetts, 1993. (Cited on page 200.)

M.F. Rushworth and T.E. Behrens. Choice, uncertainty and value in prefrontal and cingulate cortex. *Nature Neuroscience*, 11(4) :389–397, 2008. (Cited on pages 5 and 203.)

K. Samejima and K. Doya. Multiple representations of belief states and action values in corticobasal ganglia loops. *Annals of the New York Academy of Sciences*, 1104 :213–228, 2007. (Cited on pages 9 and 192.)

K. Samejima, Y. Ueda, K. Doya, and M. Kimura. Representation of action-specific reward values in the striatum. *Science*, 310 :1337–1340, 2005. (Cited on page 5.)

J. Schmidhuber, J. Zhao, and N. Schraudolph. Reinforcement learning with self-modifying policies. In *Learning to Learn*, pages 293–309. Boston : Kluwer, 1997. (Cited on page 8.)

W. Schultz. Introduction. neuroeconomics : the promise and the profit. *Philosophical Transactions of the Royal Society B*, 363 :3767–3769, 2008. (Cited on page 203.)

W. Schultz, P. Dayan, and P.R. Montague. A neural substrate of prediction and reward. *Science*, 275(5306) :1593–1599, 1997. (Cited on pages 5 and 203.)

N. Schweighofer and K. Doya. Meta-learning in reinforcement learning. *Neural Networks*, 16(1) :5–9, 2003. (Cited on page 10.)

W. Shen, M. Flajolet, P. Greengard, and D.J. Surmeier. Dichotomous dopaminergic control of striatal synaptic plasticity. *Science*, 321 :848–851, 2008. (Cited on page 5.)

B. Siciliano and O. Khatib. *Handbook of robotics*. Springer-Verlag, Berlin, 2008. (Cited on page 10.)

O. Sigaud and J. Peters. From motor learning to interaction learning in robots. In O. Sigaud and J. Peters, editors, *From Motor Learning to Interaction Learning in Robots*, pages 1–12. Springer-Verlag, publisher., 2010. (Cited on page 11.)

O. Sigaud, C. Salaun, and V. Padois. On-line regression algorithms for learning mechanical models of robots : a survey. *Robotics and Autonomous Systems*, 59(12) :1115–1129, 2011. (Cited on page 205.)

W.D. Smart and L.P. Kaelbling. Effective reinforcement learning for mobile robots. In *Proceedings of the IEEE International Conference on Robotics and Automation, ICRA'02*, pages 3404–3410. 2002. (Cited on page 10.)

R. Smith, M. Self, and P. Cheeseman. Estimating uncertain spatial relationships in robotics. In I. Cox and G. Wilfong, editors, *Autonomous Robot Vehicles*, pages 167–193. Springer-Verlag, 1990. (Cited on page 12.)

F. Stulp and O. Sigaud. Robot skill learning : From reinforcement learning to evolution strategies. *Paladyn Journal of Behavioral Robotics*, 4(1) :49–61, 2013. (Cited on page 11.)

R.S. Sutton and A.G. Barto. *Reinforcement Learning : An Introduction*. Cambridge, MA : MIT Press, 1998. (Cited on pages 4, 6, and 9.)

Y.K. Takahashi, M.R. Roesch, R.C. Wilson, K. Toreson, P. O'Donnell, Y. Niv, and G. Schoenbaum. Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. *Nature Neuroscience*, 14(12) : 1590–1597, 2011. (Cited on page 203.)

S.C. Tanaka, K. Doya, G. Okada, Y. Ueda, K.and Okamoto, and S. Yamawaki. Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature Neuroscience*, 7(8) :887–893, 2004. (Cited on page 10.)

M. A. A. van der Meer, Z. Kurth-Nelson, and A. D. Redish. Information processing in decision-making systems. *The Neuroscientist*, XX(X) :1–18, 2012. (Cited on page 192.)

M.A.A. van der Meer and A.D. Redish. Ventral striatum : a critical look at models of learning and evaluation. *Current Opinion in Neurobiology*, 21 (3) :387–392, 2011. (Cited on page 5.)

H. van Hasselt and M. Wiering. Reinforcement learning in continuous action spaces. In *IEEE Symposium on Approximate Dynamic Programming and Reinforcement Learning*, pages 272–279. 2007. (Cited on page 7.)

R. Volpe, I. Nesnas, T. Estlin, D. Mutz, R. Petras, and H. Das. The claraty architecture for robotic autonomy. In *Proceedings of the 2001 IEEE Aerospace Conference*, pages 39–46. 2001. (Cited on pages 11 and 200.)

H.H. Yin and B.J. Knowlton. The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, 7(6) :464–476, 2006. (Cited on pages 5, 6, and 195.)

H.H. Yin, S.B. Ostlund, and B.W. Balleine. Reward-guided learning beyond dopamine in the nucleus accumbens : the integrative functions of cortico-basal ganglia networks. *European Journal of Neuroscience*, 28 (8) :1437–1448, 2008. (Cited on page 5.)

Abbreviations

AC	Action Cell
ACC	Anterior Cingulate Cortex
ACh	Acetylcholine
aDMS	anterior Dorso Medial Striatum
AIC	Akaike Information Criterion
AMPA	*alpha*-Amino-3-hydroxy-5-methyl-4-isoxazolepropionic Acid
ANR	Agence Nationale de la Recherche
BG	Basal Ganglia
BIC	Bayesian Information Criterion
BIPS	Bio-Inspired Perception System (by Brain Vision Systems)
BMU	Best-Matching Unit
BOLD	Blood-oxygen-level dependent
BS	Bias System
CA1	First of the four main Hippocampus divisions
ClockS	Clockwise Search
CNRS	Centre National de la Recherche Scientifique
CO1,CO2,COn	first Correct trial, second Correct trial, ...
CONDIND	Condition Index
COR	Correct
CR	Conditioned Response
CS	Conditioned Stimulus
CRE	Conditioned Reinformcement Effect
D1	First type of Dopamine receptors
D2	Second type of Dopamine receptors
DA	Dopamine
dACC	dorsal Anterior Cingulate Cortex
DLS	Dorso Lateral Striatum
DMS	Dorso Medial Striatum
ELI	Entropy-Like Index
ENS	Ecole Normale Supérieure
FB	Feedback
FP	Fixation Point
FSCV	Fast Scan Cyclic Voltammetry
Flu	Flupentixol
FMF	Feature-Model-Free
fMRI	functional Magnetic Reasonance Imaging
GABA	Gamma-Aminobutyric Acid

GNG	Growing Neural Gas
GP	Goal-direction neuron
GQL	Generalized Q-Learning
GQLSB	Model with GQL, Shift and Bias mechanisms
GTs	Goal-Trackers
HDR	Habilitation to Direct Researches
Hpc	Hippocampus
HSD	Honest Significant Difference
IBD	Inter-Block Distance
IGs	Intermediate Group
INC	Incorrect search trials
INRIA	Institut National de la Recherche en Informatique et en Autonomatique
INSERM	Institut National de la Santé et de la Recherche Médicale
ITI	Inter-Trial Interval
ISIR	Institute of Intelligent Systems and Robotics
LAAS	Laboratory for Analysis and Architecture of Systems
LC	Landmark Cell
LiCl	Lithium Chloride
LIP6	Laboratoire d'Informatique de Paris 6 (UPMC)
LL	Log-Likelihood
LPFC	Lateral Prefrontal Cortex
LPP	Log of Posterior Probability
LSerror	Least-Square error
NAc	Nucleus Accumbens
NAcC	Nucleus Accumbens Core
NAcS	Nucleus Accumbens Shell
MB	Model-based reinforcement learning
MF	Model-free reinforcement learning
MDP	Markov Decision Process
mPFC	medial Prefrontal Cortex
NIDA	National Institute on Drug Abuse
NIH	National Institute of Health
NMDA	N-Methyl-D-Aspartate
Opt	Optimization
PC	Place Cell
PCn	Principal Component number n
PCA	Principal Component Analysis
pDMS	posterior Dorso Medial Striatum
PhD	Doctorate of Philosophy
PFC	Prefrontal Cortex
PIT	Pavlovian-Instrumental Transfer
PPn	Pedunculopontine nucleus
PRTR	Place-Recognition Triggerred Response
PS	Problem Solving task
Q	Quality of state-action pairs

QL	Q-Learning
RandS	Random Search
REP	Repetition
RL	Reinforcement Learning
ROS	Robot Operating System (by Willow Garage)
RPE	Reward Prediction Error
SBnoA	Model with Shift and Bias mechanisms but no *alpha* parameter
SBnoF	Model with Shift and Bias mechanisms but not Forgetting mechanism
SC	Signal to Change
SEA	Search
SEM	Standard Error of the Mean
SLAM	Simultaneous Localization And Mapping
SNc	Substantia Nigra pars compacta
SOM	Self-Organizing Map
S-R	Stimulus-Response
ST	Start
STscore	Statistical Test score
STD	Standard Deviation
STs	Sign-Trackers
TD	Temporal-Difference
U	Uncertainty
UCB	Upper Confidence Bound
UL	Unsupervised Learning
UN	Unpaired group
UPMC	Université Pierre et Marie Curie
US	Unconditioned Stimulus
VDF	Variance Decomposition Factor
VIF	Variation Inflation Factor
VP	Ventral Pallidum
VS	Ventral Striatum
VTA	Ventral Tegmental Area

Ce document a été préparé à l'aide de l'éditeur de texte TeXworks et du logiciel de composition typographique LaTeX 2_ε.

Title Coordination of parallel learning processes in animals and robots

Abstract This HDR manuscript presents research work at the interface between Computational Neuroscience and Cognitive Robotics aiming to better understand how animals and robots can display behavioral adaptation capabilities in their partially unknown and changing environment. Previous studies have shown that the mammalian brain combines parallel learning processes in different memory systems. During instrumental conditioning as well as navigation, this permits initial learning based on a model of the environment followed by the progressive expression of learned habits. In computational terms, this can be formalized as a progressive shift from model-based to model-free reinforcement learning. The manuscript presents : 1) Proposed computational solutions for the coordination of parallel learning processes to explain animal behavior during conditioning and navigation ; 2) Uses of learning models to analyze behavioral and neural correlates of learning ; 3) Implementations of neuro-inspired learning models in robots interacting with the real world. The manuscript highlights the gain of these exchanges between disciplines to further discuss the resulting research program.

Keywords Computational modelling, Model-based analyses of biological data, Cognitive Robotics, Reinforcement Learning, Prefrontal cortex, Basal ganglia, Dopamine

Titre Coordination de processus parallèles d'apprentissage chez les animaux et les robots

Résumé Cette thèse de HDR présente des travaux à l'interface entre Neurosciences Computationnelles et Robotique Cognitive visant à mieux comprendre comment animaux et robots peuvent faire preuve de capacités d'adaptation comportementale dans des environnements partiellement inconnus et changeants. Ils se basent sur l'observation que le cerveau coordonne différents processus parallèles d'apprentissage dans différents systèmes de mémoire. En conditionnement instrumental comme en navigation, cela permet un apprentissage initial basé sur un modèle interne du monde (model-based reinforcement learning (RL)) basculant sur l'apprentissage d'habitudes comportementales (model-free RL). Le manuscrit présente donc 1) Des solutions computationnelles proposées pour la coordination de ces processus d'apprentissage ; 2) L'utilisation de ces modèles pour l'analyses de corrélats comportementaux et neuraux de l'apprentissage ; 3) L'implémentation de ces modèles dans des robots interagissant avec le monde réel. Enfin, les échanges entre ces disciplines sont discutés dans la perspective du projet de recherche proposé.

Mots-clés Modélisation omputationelle, Analyses fondées sur un modèle de donnés biologiques, Robotique Cognitive, Apprentissage par renforcement, Cortex prefrontal, Ganglions de la base, Dopamine