Adaptive coordination of multiple learning strategies in brains and robots

Mehdi Khamassi

Sorbonne Université, CNRS, Institut des Systèmes Intelligents et de Robotique (ISIR), F-75005 Paris, France
mehdi.khamassi@sorbonne-universite.fr

Preprint of a paper to appear in Proceedings of the 9th International Conference on the Theory and Practice of Natural Computing, Springer-Verlag, 2020.

Abstract. Engineering approaches to machine learning (including robot learning) typically seek for the best learning algorithm for a particular problem, or a set problems. In contrast, the mammalian brain appears as a toolbox of different learning strategies, so that any newly encountered situation can be autonomously learned by an animal with a combination of existing learning strategies. For example, when facing a new navigation problem, a rat can either learn a map of the environment and then plan to find a path to its goal within this map. Alternatively, it can learn sequences of egocentric movements in response to identifiable features of the environment. For about 15 years, computational neuroscientists have searched for the mammalian brain's coordination mechanisms which enable it to find efficient, if not necessarily optimal, combinations of existing learning strategies to solve new problems. Understanding such coordination principles of multiple learning strategies could have great implications in robotics, to enable robots to autonomously determine which learning strategies are appropriate in different contexts. Here, we review some of the main neuroscience models for the coordination of learning strategies and present some of the early results obtained when applying these models to robot learning. We moreover highlight important energy costs which can be reduced with such bio-inspired solutions compared to current deep reinforcement learning approaches. We conclude by sketching a roadmap for further developing such bio-inspired hybrid learning approaches to robotics.

1 Introduction

The mammalian brain combines multiple learning systems whose interactions, sometimes in a competitive way, sometimes in a cooperative way, are thought to be largely responsible for the high degree of behavioral flexibility observed in mammals [1–9]. For instance, the hippocampus is a brain region playing an important role in the rapid acquisition of episodic memories – the memory of individual episodes previously experienced, such as sequences of visited places

while visiting a new city [10–12]. Together with the prefrontal cortex, the hippocampus can link these episodes so as to store in long-term memory a mental representation (or 'model') of statistical regularities of the environment [13, 9, 14]. In the spatial domain, such a mental model can take the form of a 'cognitive map' [1]. Even if it constitutes an imperfect and incomplete representation of the environment, it can be used to mentally explore the map [15, 16], or to plan potential trajectories to a desired goal before acting [17, 18]. Such a *model-based strategy* enables to rapidly and flexibly adapt to changes in goal location, since the map can be updated instantaneously with the new goal location so that the animal can plan the new correct trajectory in a one-trial learning manner [9]. Nevertheless, such a flexibility comes at the expense of time- and energy-consuming planning phases (the larger the map, the longer it takes to find the shortest path between two locations). This is typically observed when a human or an animal takes longer decision time after task changes, putatively implying a re-planning phase [17, 19].

In contrast, the basal ganglia, and especially its main input region called the striatum, is involved in the slow acquisition of procedural memories [20, 21, 7]. This type of memories is typically acquired through the repetition of sequences of egocentric movements (*e.g.,* turn left, go straight, turn right) or sequences of stimulus-triggered responses (*e.g.,* start moving in response to a flash light, then stop in response to a sound) which become behavioral 'habits' [22, 23]. These habits are known to be inflexible, resulting in slow adaptation to both changes in the environment (*e.g.,* change in goal location) and changes in motivation (*e.g.,* an overtrained rat habitually presses a food-delivering lever, even when it is satiated) [24]. Nevertheless, when these habits have been well acquired in a familiar environment, they enable the animal to make fast decisions and to perform efficient action sequences without relying on the time-consuming planning system [4, 7]. Such a learning strategy is thus called *model-free* because making a decision does not require the manipulation of an internal model to mentally represent the potential long-term consequences of the actions before acting. In contrast, it is the perception of a stimulus or the recognition of a familiar location which triggers the execution of a habitual behavioral sequence.

It is fascinating how lesion studies have highlighted some degree of modularity of the organization of learning systems within the brain. Lesioning the hippocampus impairs behavioral flexibility as well as behavioral strategies relying on a cognitive map (see [7] for a review). As a consequence, hippocampus-lesioned animals only display stimulus-response behaviors in a maze and do not seem to remember the location of previously encountered food. In contrast, animals with a lesion to what is called the dorsolateral striatum have an intact map-based behavioral strategy and perform less egocentric movements during navigation (see [7] for a review). Nevertheless, the modularity is not perfect and an important degree of distributed information processing also exists. For instance, lesions to what is called the ventral striatum seem to impair representations of reward value which are required for both model-based and model-free learning strategies (again see [7] for a review). Moreover, some brain regions do not seem

to play a specific role in learning one particular strategy, but rather a role in the coordination of these strategies. For instance, lesions to the medial prefrontal cortex only impairs the initial acquisition of model-based strategies, but not their later on expression [25]. Indeed, it seems that the medial prefrontal cortex plays a central role in the arbitration between model-based and model-free strategies [26]. As a consequence, lesioning a subpart of the medial prefrontal cortex can even restore flexible model-based strategies in overtrained rats [27].

This paper is particularly aimed at illustrating how neuroscience studies of decision-making have progressively helped understanding (and are still currently investigating) the neural mechanisms underlying animals' ability to adaptively coordinate model-based and model-free learning, and to illustrate how this biological knowledge can help towards developing behaviorally flexible autonomous robots. Since the computational models of these processes largely rely on the reinforcement learning theoretical framework, the next section will first describe the employed formalism. Then the third section will briefly review some of the neuroscience results which contribute to deciphering the principles of this coordination, and how this coordination was mathematically formalized by computational neuroscience models. The fourth section will then review some of the experimental tests of these principles in robotic scenarios. We will finally discuss the perspectives of this field of research, and how it could not only contribute to improving robots' behavioral flexibility, but also to reducing the computational cost of machine learning algorithms for robots (by enabling to skip model-based strategies when the robot autonomously recognizes that model-free strategies are sufficient).

2 Formalism adopted to describe model-based and model-free reinforcement learning

For simplicity, existing computational models are most often framed within standard Markov Decision Problem (MDP) settings, where an agent visits discrete states $s \in \mathcal{S}$, using a finite set of discrete actions $a \in \mathcal{A}$. They can encounter reward scalar values $r \in \mathbb{R}$ after performing some actions a in some states s, which they have to discover. And there is a transition probability function $\mathcal{T}(s, a, s') : (\mathcal{S}, \mathcal{A}, \mathcal{S}) \rightarrow [0; 1]$, which is a generative model underlying the statistics of the task that the agent will face, and which basically determines what is the probability of ending up in a state s' after performing an action a in state s.

In navigation scenarios, states represent unique locations in space. In neuroscience modeling studies, these states are usually equally spaced on a square grid, an information expected to be provided by place cell activity in the hippocampus, which are neurons that participate in the estimation of the animal's current location within the map. In the robot navigation experiments that will be presented later on, the state space discretization process is autonomously performed by the robot after an initial exploration of the environment. As a consequence, different states have different sizes and are unevenly distributed. It is important to note that these models can easily be extended to more distributed or con-

tinuous state representations [28], for instance when facing Partially Observable Markov Decision Process (POMDP) settings [29]. The models reviewed here can also be generalized to more continuous representations of space and actions (*e.g.,* [30, 31]). Nevertheless, we stick here to the discrete state representation for the sake of simplicity and because it has a high explanatory power.

As classically assumed within the Reinforcement Learning (RL) theoretical framework, the agent's goal is here considered to be the maximization of the expected value of the long-term reward, that is the maximization of the discounted sum of future rewards over a potentially infinite horizon: $\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$, where γ ($\gamma < 1$) is a discount factor which basically assigns a weaker weights to long-term rewards than to short-term rewards. In order to meet this objective, the agent will learn a state-action value function $Q : (\mathcal{S}, \mathcal{A}) \rightarrow \mathbb{R}$ which evaluates the total discounted sum of future rewards that the agent expects to receive when starting from a given state s, taking the action a and then following a certain (eventually learned) behavioral policy π:

$$Q^{\pi}(s,a) = \mathbb{E}\left[\sum_{t} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a, a_t = \pi(s_t), s_{t+1} \sim \mathcal{T}(s_t, a_t, .)\right] \tag{1}$$

Saying that the agent adopts a *model-based* RL strategy means that the agent will progressively try to estimate an internal model of its environment. Conventionally, this model is the combination of the estimated transition function $\hat{\mathcal{T}}(s, a, s')$ and the estimated reward function $\hat{\mathcal{R}}(s, a)$ that aims at capturing the rules that the human experimenter chooses to determine which (state,action) couples yield reward in the considered task.

Various ways to learn these two functions exist. Here, we will simply consider that the agent estimates the frequencies of occurrence of states, actions and rewards from its observations. Then, the learned internal model can be used by the agent to infer the value of each action in each possible state. This inference can be a computationally costly process, especially when all (s, a) are to be visited multiple times before reaching an accurate estimation of the state-action value function Q. Some heuristics exist to simplify the computations, or to make it less costly, like *trajectory sampling* [32] or *prioritized sweeping* [33, 34], which we review in [16]. Some alternatives to full model-based strategies exist, like the *successor representation* [35], which provides the agent with more flexibility and generalization ability than a pure model-free strategy, a smaller computational cost than a pure model-based strategy, and at the same time can contribute to describe some neural learning mechanisms in the hippocampus [36, 37]. Nevertheless, for the sake of simplicity, here we will consider that the inference process in the model-based (MB) RL agent is performed through a *value iteration* process [32]:

$$Q_{MB}^{(t+1)}(s,a) = \hat{\mathcal{R}}^{(t)}(s,a) + \gamma \sum_{s'} \hat{\mathcal{T}}^{(t)}(s,a,s') \max_{k \in \mathcal{A}} Q_{MB}^{(t)}(s',k) \tag{2}$$

In contrast, an agent adopting a *model-free* RL strategy will not have access to nor try to estimate any internal model of the environment. Instead, the agent

will iteratively update its estimation of state-action value function through its interactions with the environment. Each of these interactions consist in performing an action a in a state s and observing the consequence in terms of the reward r that this yields and the new state s' of the environment. Again, if we address a navigation problem, the possible actions are typically movements towards cardinal directions (North, South, East, West) and the new state s' is the new location of the agent within the environment after acting. A classical and widely used model-free RL algorithm is Q-learning [38]:

$$Q_{MF}^{(t+1)}(s_t, a_t) = Q_{MF}^{(t)}(s_t, a_t) + \alpha(r_t + \gamma \max_{k \in \mathcal{A}} Q_{MF}^{(t)}(s_{t+1}, k) - Q_{MF}^{(t)}(s_t, a_t)) \quad (3)$$

where $\alpha \in [0; 1]$ is the learning rate, and the term between parentheses, often written δ_t is called the *temporal-difference error* [32] or the *reward prediction error* [39] because it constitutes a reinforcement signal which compares the new estimation of value $(r_t + \gamma \max_{k \in \mathcal{A}} Q_{MF}^{(t)}(s_{t+1}, k))$ after performing action a_t in state s_t, arriving in state s_{t+1} and receiving a reward r_t, with the expected value $Q_{MF}^{(t)}(s_t, a_t)$ before executing this action. Any deviation between the two is used as an error signal to correct the current estimation of the state-action value function Q.

Finally, for the decision-making phase, no matter if the agent is model-free or model-based, the agent selects the next action a to perform from a probability distribution over actions in the current state s computed from the estimated state-action value function $x^{(t)}(s, a)$ (with $x^{(t)}(s, a) = Q_{MB}^{(t)}(s, a)$ if the agent is model-based, or $x^{(t)}(s, a) = Q_{MF}^{(t)}(s, a)$ if the agent is model-free), using a Boltzmann softmax function:

$$P^{(t)}(a|s) = \frac{e^{\beta x^{(t)}(s,a)}}{\sum_{k \in \mathcal{A}} e^{\beta x^{(t)}(s,k)}} \quad (4)$$

where β is called the *inverse temperature* which regulates the exploration/exploitation trade-off by modulating the level of stochasticity of choice: the closer β is to zero, the more the contrast between Q-values will be attenuated, the extreme being for $\beta = 0$ which produces a flat action probability distribution (random exploration); in contrast, the larger the value of β, the more the contrast between Q-values will be enhanced, which makes the probability of the action with the highest Q-value close to 1 when β tends towards ∞ (exploitation).

3 Neuroscience studies of the coordination of model-based and model-free reinforcement learning

Reinforcement learning models (initially only from the model-free family) have started to become popular in neuroscience in the mid 90s, when researchers discovered that a part of the brain called the dopaminergic system (because it innervates the rest of the brain with a neuromodulator called *dopamine*) increases its activity in response to unpredicted reward, decreases its activity in

response to the absence of a predicted reward, but does not respond to predicted ones, as can be modeled when δ_t, the right part between parentheses in Equation 3, is positive, negative or null, respectively [39]. This discovery was followed by a large set of diverse neuroscience experiments to verify that other parts of the brain could show neural activity compatible with other variables of reinforcement learning models like state-action values (see examples of comparisons of neuroscience results with RL models' predictions in [40, 41]; and see [5] for a review).

More important for the topic of this paper, since about 10 years, an increasing number of neuroscience studies have started to investigate whether human and animal behavior during reward-based learning tasks could involve some sort of combination of model-based (MB) and model-free (MF) learning processes, and what are the neural mechanisms underlying such a coordination.

The simplest possible way of combining MB and MF RL processes is to consider that they occur in parallel in the brain, and that any decision made by the subject results from the simple weighted sum of MB and MF state-action values (*i.e.,* replacing $x^{(t)}(s,a)$ in Equation 4 by $(1-\omega)Q^{(t)}_{MB}(s_t,a_t) + \omega Q^{(t)}_{MF}(s_t,a_t)$, with $\omega \in [0;1]$ a weighting parameter). This works well to have a first approximation of the degree with which different individual subjects, whether human adults [42], human children [43], or animals [44], rely on a model-based system to make decisions while considering the ensemble of trials made by the subject during a whole experiment. This has for instance helped understand that children make more model-free decisions than adults because their brain area subserving model-based decisions (their prefrontal cortex) takes years to mature [43]. This has also helped better model why some subjects are more prone than others to be influenced by reward predicting stimuli (which has implication to understand stimulus-triggered drug-taking behaviors in humans): roughly because their model-based process contributes less to their decisions [44].

Nevertheless, a systematic weighted sum of MB's and MF's decisions has the disadvantage of systematically requiring the (potentially heavy) computations from both learning systems. In contrast, it is thought that relying on habitual behaviors learned by the model-free system when the environment is stable and familiar is useful to avoid the costly inference-related computations of the model-based system [4, 5]. There could thus be some evolutionary reasons why humans do not always perform rational choices as could be (more often) the case if they were relying more on their model-based system [45]: namely that they would not be able to make fast decisions in easy familiar situations. Instead, they would always need to make long inferences with their internal models before deciding. Thus, they would be exhausted at the end of the day if they had to think deeply for all the simple decisions they have to make everyday, like whether they should drink a coffee before taking a shower or the opposite, whether they should wear a blue or a red shirt, where to go for lunch, etc.

Alternatively, early neuroscience studies of the coordination of MB and MF process hypothesized a sequential activation of the two systems: humans and animals should initially rely on their MB system when facing a new task, so

as to figure out what are the statistics of the task and what is the optimal thing to do; and as they repeat over and over the same task and get habituated to it (making it become familiar), they should switch to the less costly MF system which hopefully will have had time to learn during a long repetition phase. Moreover, if suddenly the task changes, they should restart using their MB system (and thus break their previously acquired habit) in order to figure out what has changed and what is the new optimal behavioral policy. And then again after many repetitions with the new task settings, they can acquire a new behavioral habit with the MF system.

An illustrative example is the case where someone has to visit a new city. In that case, people usually look at a map, which is an allocentric representation of the city, and try to remember the parts of the maps that are useful to reach a desired location. And then, once people walk through the city, if they suddenly find themselves in front of some landmark that they thought would not be encountered during their planned trip (*e.g.,* a monument), they can close their eyes and try and understand where they might actually be located within their mental map, and which change in their trajectory they should operate. This is typically a MB inference process. In contrast, when one always takes the same path from their home until their workplace, they rarely perform MB inference, and rather let their body automatically turn at the right corner and lead them to their usual arrival point. This works well even if one is discussing with a friend while walking, or is not fully awake. We thus think that in such a case the brain has shifted its decisions to the MF system. This permits to free other parts of the brain which can be used to think while we walk about the last book we read, or to try and solve the maths problem we are currently addressing.

One initial computational proposal for the coordination of MB and MF learning systems which can well capture this dynamics consists in comparing the uncertainty of the MB and MF systems and relying on the most certain one [4]. This can be achieved if a Bayesian formulation of RL is adopted where the agent does not simply learns point estimates of state-action value functions, but rather full distributions over each (state,action) pair value. In that case, the precision of the distribution can be used to represent the level of uncertainty. In practice, when facing a new task, the uncertainty in both systems is high. But the uncertainty in the MB system decreases faster with learning (*i.e.,* after less observations made following interactions of the agent with the world, even if these observations are processed during a long inference phase). As a consequence, the agent will rely more on the MB system during early learning. In parallel the uncertainty in the MF system slowly decreases, until the MF system is sufficiently certain to take control over the agent's actions. When the task changes (*e.g.,* the goal location changes, or a path is now obstructed by an obstacle), uncertainty re-increases in both systems, but again it decreases faster in the MB system, so that again a sequence of MB decisions followed by MF decisions after a long second learning phase can be produced.

However, systematically monitoring uncertainty in the MB system can be computational heavy, and does not really permit the avoidance of the costly

computations of the MB system when the MF system is currently leading. Alternatively, a more recent computational neuroscience model proposes to only monitor uncertainty within the MF system, and considers that the MB system is by default providing *perfect information*, so that it should be chosen when the MF system is uncertain, and avoided only when the MF system is sufficiently certain [6]. This works well in a number of situations and enables to well capture the behavior of animals in a number of experiments. However, there are situations where this assumption cannot be true. In particular, as we will illustrate with some robotic tests of this kind of models in the next section, if the agent has an inaccurate internal model, it is better to rely on the more reactive MF system even when it is still uncertain [46, 47]. Less costly alternative estimations of uncertainty can be used to permanently monitor both the MB and the MF system. For instance, the degree of instability of Q-values (MB or MF) before convergence is reached can be a good proxy to uncertainty [16, 48]. Choice confidence can also give a relatively good proxy to choice uncertainty in simple situations, by measuring the entropy of the probability distribution over all possible actions in a given state and comparing MB and MF estimations of this measure [19]. This moreover enables to well capture not only choices made by human subjects during simple decision-making tasks, but also their decision time (the more uncertain they are, the more time they need to make their decision). Finally, this type of mechanism also enables to explain why the ideal MB-to-MF sequence is not always true, since early choices of human subjects can sometimes significantly rely on the MF system because their MF system might initially be overconfident [19].

Another important current question is whether uncertainty alone is sufficient to arbitrate between MB and MF systems [49], or whether, when the two systems are equally uncertain, the agent should rely on the least computationally costly one [50]. If we want the agent to be initially agnostic about which system is more costly, and if we even want the agent to be able to potentially arbitrate between N different learning systems with different a priori unknown computational characteristics, then one proposal is simply to measure the average time taken by each system when it has to make decisions [50]. In some of the robotic experiments that we will describe in the next section, we found that this principle works robustly, enables to produce the ideal MB-to-MF sequence, not only during initial learning but also after a task change. We will come back to this later.

Finally, other current outstanding questions are whether the two systems shall always be in competition, or whether they shall sometimes also cooperate (as can be achieved with the weighted sum of their contribution described above); and whether an efficient coordination mechanism shall arbitrate between MB and MF at each timestep from the current available measures (*e.g.,* uncertainty, computational cost, etc.), or whether it is sometimes more efficient to learn and remember that the MF system is usually better in situation type A while the MB system is better in situation type B. The latter could enable the agent to instantaneously rely on the best memorized system without needing

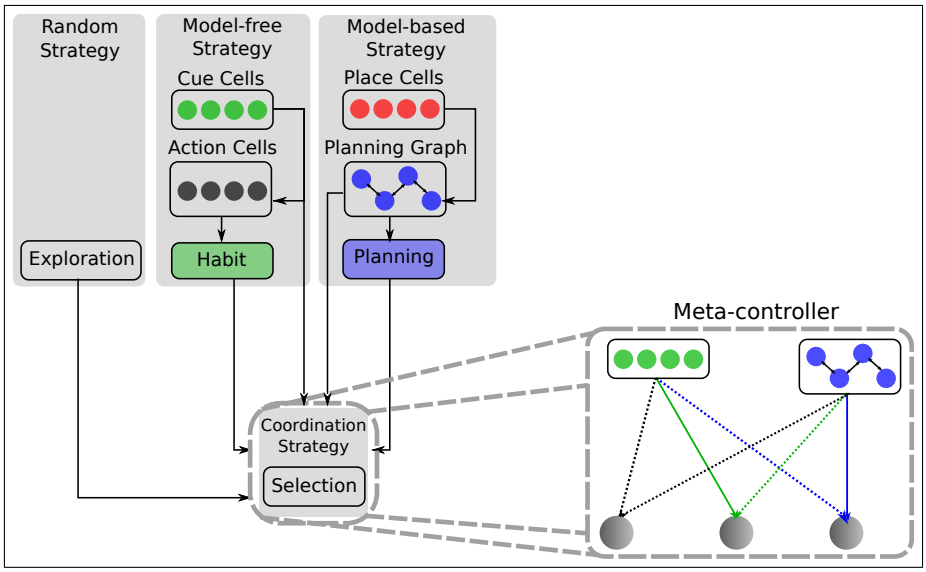

Fig. 1. The generic architecture for the coordination of multiple learning strategies applied to navigation proposed by Dollé and colleagues [9]. Three main learning strategies are considered here (but the paper tests other variants, such as the combination of multiple instances of the same strategy, which would correspond to a case of ensemble learning [51]): a model-based planning strategy; a model-free habitual strategy; and a distinct random exploration strategy in order to avoid cumulating the exploratory decisions of the two other strategies. A so-called 'meta-controller' performs the high-level strategy coordination process. This process can consists in different manners of coordinating strategy, such as giving the hand to the least uncertain one [4]. Nevertheless, in [9] the meta-controller learns which strategy yields the largest amount of reward in different locations of the environment or in the presence of different stimuli. Adapted from [9].

to fully experience a new situation identified as belonging from a recognized type. This issue relates to current investigations in subfields of machine learning known as transfer learning, life-long learning and open-ended learning [52–54]. One solution to this coordination memory problem consists in adopting a hierarchical organization where a second, higher-level, learning process (in what Dollé and colleagues call a *'meta-controller'*) learns which strategy (model-based, model-free or random exploration) is the best (in terms of the amount of reward it yields) in different parts of the environment [9] (Fig. 1). This model learns through RL which strategy is the most efficient in each part of the environment. It can moreover learn that a certain equilibrium between MB and MF processes is required for good performance, thus resulting in cooperation between systems. It can even learn to change through time the weight of the contribution of each system, as learning in the MF system progresses, thus producing something that looks like the ideal MB-to-MF sequence.

With these principles in hand, the Dollé model [9] can explain a variety of rat navigation behaviors experimentally observed, including data that initially appeared as either contradictory to the cognitive map theory, or contradictory to the associative learning theory which approximately considers that navigation behaviors shall all be learned through model-free reinforcement learning. Finally, it is worthy of note that performing offline replay of the MB system can result in learning by observation in the MF system, so that the two somehow cooperate [16], as inspired by the now classical DYNA architecture [55].

Overall, this short review highlights that the investigation of the principles underlying the adaptive coordination of model-based and model-free reinforcement learning mechanisms in humans and animals is currently an active area of research in neuroscience.

4 Robotic tests of bio-inspired principles for the coordination of model-based and model-free reinforcement learning

The importation of these bio-inspired ideas to robotics is quite recent, and is still an emerging field of research. Nevertheless, a few studies and their outcomes deserve to be mentioned here.

To our knowledge, the first test with a real robot of a bio-inspired algorithm for the online coordination of model-based and model-free reinforcement learning has been presented in [46]. This work included an indoor robot navigation scenario with the Psikharpax rat robot [56] within a 2mx2.5m arena (Fig. 2). The robot first explored the environment to autonomously learn a cognitive map of the environment (hence a mental model used by its model-based learning strategy). In addition, the robot could use a model-free reinforcement learning strategy to learn movements in 8 cardinal directions in response to perceived salient features within the environment (*i.e., stimuli* in the vocabulary of psychology). The latter MF RL component of the model was later improved in [57] to make it able to learn movements away from visual features when needed. The proposed algorithm for the online coordination of MB and MF RL was based on the computational neuroscience model of Dollé and colleagues [58, 59, 9], which has been presented in the previous section and sketched in Fig. 1.

The first important result of this robotic work is that the algorithm could autonomously learn the appropriate coordination of MB and MF systems for each specific configuration of the environment that was presented to the robot. For a first configuration associated to an initial goal location (where the robot can obtain a scalar reward), the algorithm learned that the MB strategy was appropriate to guide the robot from far away towards the goal, and that the MF strategy was appropriate to control the fine movements of the robot when closer to the goal. This was an emergent property of the coordination that was not designed by human. Instead, it was autonomously learned by the algorithm in response to the specific environment where the robot was located. The reason is that the robot had less explored the area around the initial goal location.

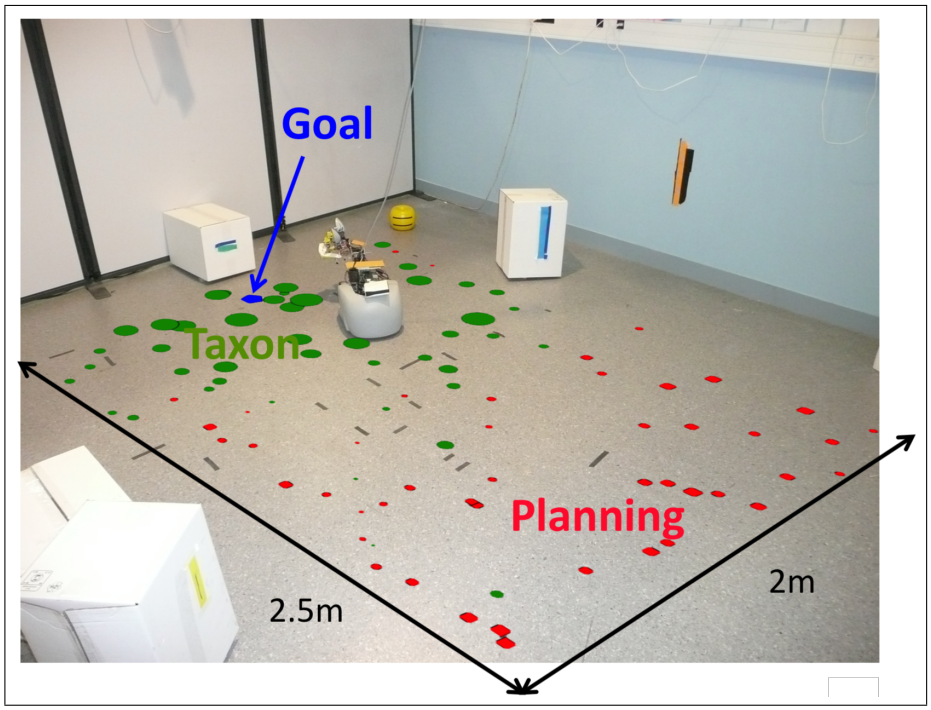

Fig. 2. Robotic experiments presented in [46] and aiming at testing the performance of a bio-inspired algorithm for the online coordination of model-based and model-free reinforcement learning. The algorithm itself is based on the computational neuroscience model of Dollé and colleagues [9], which is presented in Fig. 1. The photo shows the Psikharpax rat robot [56] within a 2mx2.5m indoor arena. Salient features are displayed on the surrounding walls and objects because the purpose of this research work was not on vision, but how to make the robot use the noisy perception of these features with its cameras in order to learn a simple cognitive map that can be used for model-based navigation. The green and red dots superimposed on the photo show the location of the *place cells* of the cognitive map learned by the robot. The blue dot (invisible to the robot) shows the current location of the goal, where the robot receives a scalar reward during the first part of the experiment. Later on this goal location is moved without informing the robot, so that it needs to detect this change and learn a new appropriate coordination of model-based and model-free strategies to quickly reach a good performance. In the last part of the experiment, the goal is again moved back to its initial location, to show that the coordination algorithm can, after detecting this new change, instantaneously retrieve its memory of the first appropriate coordination and thus quickly re-display a good performance without re-learning. Finally, the green dots are labelled 'taxon' (a *taxon* strategy in neuroscience consists in learning actions in response to visual cues), which corresponds to areas where the algorithm considers the model-free strategy as the best strategy, while red dots are labelled 'planning' because the algorithm considers that the model-based strategy is the most appropriate there. Adapted from [46].

Thus, its cognitive map was less precise there. As a consequence, a pure MB version of the algorithm could learn to approach the robot near the goal, but could not learn to precisely reach it (because of the imprecision in the map). As a consequence, the autonomous coordination algorithm found out that the MF system could compensate for this lack of precision. From this simple example we can learn two things: first, that in contradiction to the assumption made by some previously discussed computational models that the MB system has access to *perfect information*, the map (*i.e.,* model) learned by the MB system can be imperfect, and the coordination algorithm has to cope with it. More generally, we think that when experimenting with robots, there will always be a situation where the map cannot be accurately learned, because of noisy perceptions, problems with the light, etc. So, rather than endlessly trying to refine the MB system to make it appropriate for each new situation at hand, it might be better to let the coordination algorithm autonomously find out what is the appropriate alternation between MB and MF for the present situation. The second thing that we can learn from this example is that a simple coordination algorithm which puts MB and MF systems in competition, and selects the most efficient one, can sometimes produce a sort of cooperation between them. In this particular example, a learned trajectory of the robot to the goal can be hybrid, involving a first use of the MB strategy when far away from the goal, and then a shift to the MF strategy when getting closer. This enables us to draw a model-driven prediction for neuroscience: that sometimes animals solving these types of task may display a trajectory within a maze or an arena that is not the result of a single learning system, but rather a hybrid byproduct of the coordination of multiple systems.

Another important result of this robotic work relates to its ability to learn context-specific coordination patterns, which can relate to what people call *episodic control*. This occurred when we changed the goal location after some time, and let the robot adapt to the new configuration. What happened is that the algorithm first detected the change because of the different profile of reward propagation through its mental map that this induced. Then the algorithm decided to store in memory the previously learned coordination pattern between MB and MF, and to learn a new one. After a new learning phase, the algorithm found a new coordination pattern adapted to the new condition, thus producing good performance again. Finally, we suddenly moved the goal location back to its initial location. The algorithm could detect the change and recognize the previous configuration (again thanks to the profile of reward propagation through its mental map). As a consequence, the algorithm retrieved the previously learned coordination pattern, which enabled the robot to instantaneously restore the appropriate behavior without re-learning.

Nevertheless, some limitations and perspectives of this seminal work ought to be mentioned here. First, the coordination component of the algorithm (which is called the *meta-controller* in [9, 46, 57]) slowly learns through MF RL (in addition to the MF RL mechanism used within the MF system dedicated to the MF strategy) which strategy is the most appropriate in each part of the environment

(In other words, the model involves a hierarchical learning process in addition to the parallel learning process between MB and MF strategies). While this is good for the robot to be able to memorize specific coordination patterns for each context (*i.e.,* for each configuration of the goal location within the arena), this nevertheless requires a long time to achieve a good coordination within each context. Thus, it would be interesting to also test coordination mechanisms based on instantaneous measures such as uncertainty, as discussed in the previous section. A second limitation is that this experiment involved a specific adaptation of a coordination model to a simple indoor navigation task, with a small map, a small number of states to learn, and an action repertoire which is specific to navigation scenarios. A third limitation is at the technical level, involving an old custom robot. Thus, it it not clear if these results could be generalized to other robotic platforms facing a wider variety of tasks, and sometimes more complex tasks involving a larger number of states.

Fig. 3. (Left) Human-Robot Interaction task tested in [60]: the human and the robot collaborate to put all boxes in a trashbin. (Right) Navigation task autonomously mapped and discretized by the robot during exploration [60, 50]. The red area indicates the goal location whereas the green areas indicate starting locations of the robot. Red numbers are starting location indexes; blue numbers are some states where some changes in the configuration of the environment can occur. Adapted from [60].

A more recent series of robotic experiments with the same research goal (*i.e.,* assessing the efficience and robustness of bio-inspired coordination principles of MB and MF learning) has been presented in [61, 62, 47] and later in [60, 50, 63]. First, [61, 62, 47] compared different coordination principles, including methods coming from ensemble learning [51] in several different simulated robotic tasks. They found again that the MB system was not always the most reliable system, especially in tasks with hundreds of states, where the MB system requires long inference durations to come up with a good approximation of the state-action value function Q. These experiments highlighted the respective advantages and disadvantages of MB and MF reinforcement learning in a variety of simulated robotic situations, and concluded again for the added value of coordinating them. In [60, 50, 63], simulated and real robotic experiments were presented, some involving navigation with a Turtlebot, and others involving simulated tasks with the PR2 robot and the Baxter robot (Fig. 3).

The first important result of these new series of experiments to highlight is that the coordination of MB and MF RL was efficient in a variety of tasks, including navigation tasks with detours, non-stationarity of the configuration of the environment (*i.e.,* sudden introduction of obstacles obstructing some corridors), but also simple human-robot interaction tasks. In the latters, the human and the robot had to cooperate to clean a table by putting objects in a trashbin. Importantly, some objects were reachable by the human, some by the robots, thus forcing them to communicate and cooperate. In that case, the model-based system could compute joint action plans where actions by the robot and actions by the humans alternated. In all these situations, the robot could autonomously learn the task and reach good performance.

The second important result to highlight is that instantaneous measures of uncertainty in MB and MF systems allow a quicker reaction of the coordination mechanism to changes in the environment. Nevertheless, this does not permit memorization nor episodic control, which the work of [46] did. Thus, the results are in good complementarity and in the future it would interesting to test combinations of these two principles.

The last important result to highlight here is that an efficient coordination mechanism proposed by [50, 63], and successfully applied to robot navigation and human-robot interaction scenarios, consists in taking into account not only the uncertainty but also the computational cost of each learning system. In practice, the proposed algorithm monitored the average time taken by each system to make its inference phase before deciding. It learned that the MB system takes on average 10 times longer than the MF system, in these specific tasks, before making a decision. As a consequence, the coordination algorithm gave the lead to the MF system in cases of equal uncertainty, and even in cases of slightly higher uncertainty in the MF system. As a result, the algorithm mostly relied on the MF system but transiently and efficiently gave the lead to the MB system, only when needed. This occurred during initial learning as well as after task changes. As a consequence, the algorithm could reproduce the nice MB-to-MF sequence that we discussed in previous sections, both during initial learning and after task changes. Moreover, with this new coordination principle, the robot could achieve the same optimal performance as a pure MB system (which was optimal in these cases) while requiring a cumulated computational cost which was closer to that of a pure MF system (which achieves a lower bound on computational cost in these experiments). Thus, the coordination algorithm not only allowed for an efficient and flexible behavior of the robot in these non-stationary tasks, but it also permitted to reduce the computational cost of the algorithm controlling the robot. Finally, the authors also compared their algorithm with a state-of-the-art deep reinforcement learning algorithm. They found that the latter requires a very large number of iterations to learn, much more than their proposed MB-MF coordination algorithm.

Finally, it is interesting to mention that in the meantime, several other research groups throughout the world have also started to test hybrid MB/MF algorithms for robot learning applications [64–69]. In particular, the deep re-

inforcement learning community is showing a growing interest for such hybrid learning algorithms [70–72]. This illustrates the potentially broad interest that this type of hybrid solutions to reinforcement learning can have in different research communities.

5 Conclusion

This paper aimed at first illustrating current outstanding questions and investigations to better understand and model neural mechanisms for the online adaptive coordination of multiple learning strategies in humans and animals. Secondly, the paper reviewed a series of recent robot learning experiments aimed at testing such bio-inspired principles for the coordination of model-based and model-free reinforcement learning strategies.

We discussed the respective advantages and disadvantages of different coordination mechanisms: on the one hand, mechanisms relying on instantaneous measures of uncertainty, choice confidence, performance, as well as computational cost; on the other hand, mechanisms relying on hierarchical learning where a high-level *meta-controller* autonomously learns which strategy is the most efficient in each situation.

The robotic experiments discussed here showed that this type of coordination principle can work efficiently, robustly and at a reduced computational cost in a variety of robotic scenarios (navigation, human-robot interaction). This is of particular importance at a time where energy saving is a critical issue for the planet and to slow down global warming. In contrast, many current machine learning techniques, especially those relying on deep learning, require tremendous amounts of energy and long pre-training phases.

Finally, the paper aimed at also illustrating the interest of testing neuro-inspired models in real robots interacting with the real world so as to generate novel model-driven predictions for neuroscience and psychology. In the particular case of the adaptive coordination of model-based and model-free reinforcement learning strategies, we showed that some situations can induce cooperation between learning strategies. We moreover showed that not only taking into account the uncertainty of each learning system but also its computational cost could work efficiently in a variety of task. This raises the prediction that the mammalian brain may also monitor and memorize the average computational cost (for instance in terms of the duration required for inference) of different learning strategies in different memory systems, in order to favor those which cost less when they are equally efficient. This paves the way for novel neuroscience experiments aimed at testing these new model-driven predictions and understanding the underlying neural mechanisms.

Disclosure/Conflict-of-Interest Statement

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments

The author would like to thank all his collaborators who have contributed through the years to this line of research. In particular, Andrea Brovelli, Romain Cazé, Ricardo Chavarriaga, Laurent Dollé, Benoît Girard, Agnes Guillot, Mark Humphries, Florian Lesaint, Olivier Sigaud, Guillaume Viejo for their contribution to the design, implementation, test, and analysis of computational models of the coordination of learning processes in humans and animals. And Rachid Alami, Lise Aubin, Ken Caluwaerts, Raja Chatila, Aurélie Clodic, Sandra Devin, Rémi Dromnelle, Antoine Favre-Félix, Benoît Girard, Christophe Grand, Agnes Guillot, Jean-Arcady Meyer, Steve N'Guyen, Guillaume Pourcel, Erwan Renaudo, Mariacarla Staffa for their contribution to the design, implementation, test and analysis of robotic experiments aimed at testing neuro-inspired principles for the coordination of learning processes.

Funding

This work has been funded by the Centre National de la Recherche Scientifique (CNRS)'s interdisciplinary programs (MITI) under the grant name 'Hippocampal replay through the prism of reinforcement learning'.

References

1. John O'keefe and Lynn Nadel. *The hippocampus as a cognitive map*. Oxford: Clarendon Press, 1978.
2. James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995.
3. Hiroyuki Nakahara, Kenji Doya, and Okihide Hikosaka. Parallel cortico-basal ganglia mechanisms for acquisition and execution of visuomotor sequences-a computational approach. *Journal of cognitive neuroscience*, 13(5):626–647, 2001.
4. Nathaniel D Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, 8(12):1704–1711, 2005.
5. Mehdi 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-Paris VI, 2007.

6. Mehdi Keramati, Amir Dezfouli, and Payam Piray. Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Comput Biol*, 7(5):e1002055, 2011.
7. Mehdi Khamassi and Mark D Humphries. Integrating cortico-limbic-basal ganglia architectures for learning model-based and model-free navigation strategies. *Frontiers in behavioral neuroscience*, 6:79, 2012.
8. Giovanni Pezzulo, Francesco Rigoli, and Fabian Chersi. The mixed instrumental controller: using value of information to combine habitual choice and mental simulation. *Frontiers in psychology*, 4:92, 2013.
9. Laurent Dollé, Ricardo Chavarriaga, Agnès Guillot, and Mehdi Khamassi. Interactions of spatial strategies producing generalization gradient and blocking: A computational approach. *PLoS computational biology*, 14(4):e1006092, 2018.
10. Neil Burgess, Eleanor A Maguire, and John O'Keefe. The human hippocampus and spatial and episodic memory. *Neuron*, 35(4):625–641, 2002.
11. Stefan Leutgeb, Jill K Leutgeb, Carol A Barnes, Edvard I Moser, Bruce L McNaughton, and May-Britt Moser. Independent codes for spatial and episodic memory in hippocampal neuronal ensembles. *Science*, 309(5734):619–623, 2005.
12. Howard Eichenbaum. Prefrontal–hippocampal interactions in episodic memory. *Nature Reviews Neuroscience*, 18(9):547–558, 2017.
13. Paul W Frankland and Bruno Bontempi. The organization of recent and remote memories. *Nature reviews neuroscience*, 6(2):119–130, 2005.
14. Ivilin Stoianov, Domenico Maisto, and Giovanni Pezzulo. The hippocampal formation as a hierarchical generative model supporting generative replay and continual learning. *bioRxiv*, 2020.
15. Anoopum S Gupta, Matthijs AA van der Meer, David S Touretzky, and A David Redish. Hippocampal replay is not a simple function of experience. *Neuron*, 65(5):695–705, 2010.
16. Romain Cazé, Mehdi Khamassi, Lise Aubin, and Benoît Girard. Hippocampal replays under the scrutiny of reinforcement learning models. *Journal of neurophysiology*, 120(6):2877–2896, 2018.
17. Adam Johnson and A David Redish. Neural ensembles in ca3 transiently encode paths forward of the animal at a decision point. *Journal of Neuroscience*, 27(45):12176–12189, 2007.
18. Marcelo G Mattar and Nathaniel D Daw. Prioritized memory access explains planning and hippocampal replay. *Nature neuroscience*, 21(11):1609–1617, 2018.
19. Guillaume Viejo, Mehdi Khamassi, Andrea Brovelli, and Benoît Girard. Modeling choice and reaction time during arbitrary visuomotor learning through the coordination of adaptive working memory and reinforcement learning. *Frontiers in behavioral neuroscience*, 9:225, 2015.
20. Steven P Wise. The role of the basal ganglia in procedural memory. In *Seminars in Neuroscience*, volume 8, pages 39–46. Elsevier, 1996.
21. Mark G Packard and Barbara J Knowlton. Learning and memory functions of the basal ganglia. *Annual review of neuroscience*, 25(1):563–593, 2002.
22. Mandar S Jog, Yasuo Kubota, Christopher I Connolly, Viveka Hillegaart, and Ann M Graybiel. Building neural representations of habits. *Science*, 286(5445):1745–1749, 1999.
23. Amir Dezfouli and Bernard W Balleine. Habits, action sequences and reinforcement learning. *European Journal of Neuroscience*, 35(7):1036–1051, 2012.
24. Anthony Dickinson and Bernard Balleine. Motivational control of goal-directed action. *Animal Learning & Behavior*, 22(1):1–18, 1994.

25. Sean B Ostlund and Bernard W Balleine. Lesions of medial prefrontal cortex disrupt the acquisition but not the expression of goal-directed learning. *Journal of Neuroscience*, 25(34):7763–7770, 2005.
26. Simon Killcross and Etienne Coutureau. Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral cortex*, 13(4):400–408, 2003.
27. Etienne Coutureau and Simon Killcross. Inactivation of the infralimbic prefrontal cortex reinstates goal-directed responding in overtrained rats. *Behavioural brain research*, 146(1-2):167–174, 2003.
28. 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(3):639–652, May 2004.
29. Anthony R Cassandra, Leslie Pack Kaelbling, and Michael L Littman. Acting optimally in partially observable stochastic domains. In *Aaai*, volume 94, pages 1023–1028, 1994.
30. Kenji Doya. Reinforcement learning in continuous time and space. *Neural computation*, 12(1):219–245, 2000.
31. Mehdi Khamassi, George Velentzas, Theodore Tsitsimis, and Costas Tzafestas. Robot fast adaptation to changes in human engagement during simulated dynamic social interaction with active exploration in parameterized reinforcement learning. *IEEE Transactions on Cognitive and Developmental Systems*, 10(4):881–893, 2018.
32. Richard S. Sutton and Andrew G. Barto. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998.
33. Jing Peng and Ronald J Williams. Efficient learning and planning within the Dyna framework. *Adaptive Behavior*, 1(4):437–454, 1993.
34. Andrew W Moore and Christopher G Atkeson. Prioritized sweeping: Reinforcement learning with less data and less time. *Machine learning*, 13(1):103–130, 1993.
35. Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624, 1993.
36. Kimberly L Stachenfeld, Matthew M Botvinick, and Samuel J Gershman. The hippocampus as a predictive map. *Nature neuroscience*, 20(11):1643, 2017.
37. Ida Momennejad. Learning structures: Predictive representations, replay, and generalization. *Current Opinion in Behavioral Sciences*, 32:155–166, 2020.
38. Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.
39. W. Schultz, P. Dayan, and P. R. Montague. A neural substrate of prediction and reward. *Science*, 275:1593–1599, 1997.
40. Jean Bellot, Olivier Sigaud, and Mehdi Khamassi. Which temporal difference learning algorithm best reproduces dopamine activity in a multi-choice task? In *International Conference on Simulation of Adaptive Behavior*, pages 289–298. Springer, 2012.
41. Jean Bellot, Olivier Sigaud, Matthew R Roesch, Geoffrey Schoenbaum, Benoît Girard, and Mehdi Khamassi. Dopamine neurons activity in a multi-choice task: reward prediction error or value function? In *Proceedings of the French Computational Neuroscience NeuroComp12 workshop*, pages 1–7, 2012.
42. Nathaniel D Daw, Samuel J Gershman, Ben Seymour, Peter Dayan, and Raymond J Dolan. Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204–1215, 2011.
43. Johannes H Decker, A Ross Otto, Nathaniel D Daw, and Catherine A Hartley. From creatures of habit to goal-directed learners: Tracking the developmental emergence of model-based reinforcement learning. *Psychological science*, 27(6):848–858, 2016.

44. 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 Comp. Biol.*, 10(2), feb 2014.
45. Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.
46. 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*, 7:025009, 2012.
47. Erwan Renaudo, Benoît Girard, Raja Chatila, and Mehdi Khamassi. Respective advantages and disadvantages of model-based and model-free reinforcement learning in a robotics neuro-inspired cognitive architecture. In *Biologically Inspired Cognitive Architectures BICA 2015*, pages 178–184, Lyon, France, 2015.
48. Mehdi Khamassi and Benoît Girard. Modeling awake hippocampal reactivations with model-based bidirectional search. *Biological Cybernetics*, pages 1–18, 2020.
49. John Philip O'Doherty, Sangwan Lee, Reza Tadayonnejad, Jeff Cockburn, Kiyohito Iigaya, and Caroline J Charpentier. Why and how the brain weights contributions from a mixture of experts. 2020.
50. R. Dromnelle, E. Renaudo, G. Pourcel, R. Chatila, B. Girard, and M. Khamassi. How to reduce computation time while sparing performance during robot navigation? a neuro-inspired architecture for autonomous shifting between model-based and model-free learning. In *9th International conference on biomimetic & biohybrid systems (Living Machines 2020)*, LNAI, pages 1–12, Online conference (initially planned in Freiburg, Germany), 2020.
51. Marco A. Wiering and Hado van Hasselt. Ensemble algorithms in reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 38(4):930–936, 2008.
52. Sebastian Thrun. Lifelong learning algorithms. In *Learning to learn*, pages 181–209. Springer, 1998.
53. Paul Ruvolo and Eric Eaton. Ella: An efficient lifelong learning algorithm. In *International Conference on Machine Learning*, pages 507–515, 2013.
54. Stephane Doncieux, Nicolas Bredeche, Léni Le Goff, Benoît Girard, Alexandre Coninx, Olivier Sigaud, Mehdi Khamassi, Natalia Díaz-Rodríguez, David Filliat, Timothy Hospedales, et al. Dream architecture: a developmental approach to open-ended learning in robotics. *arXiv preprint arXiv:2005.06223*, 2020.
55. Richard S Sutton. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Proceedings of the seventh international conference on machine learning*, pages 216–224, 1990.
56. Jean-Arcady Meyer, Agnès Guillot, Benoît Girard, Mehdi Khamassi, Patrick Pirim, and Alain Berthoz. The psikharpax project: Towards building an artificial rat. *Robotics and autonomous systems*, 50(4):211–223, 2005.
57. 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.
58. Laurent Dollé, Mehdi Khamassi, Benoît Girard, Agnes Guillot, and Ricardo Chavarriaga. Analyzing interactions between navigation strategies using a computational model of action selection. In *International Conference on Spatial Cognition*, pages 71–86. Springer, 2008.
59. Laurent Dollé, Denis Sheynikhovich, Benoît Girard, Ricardo Chavarriaga, and Agnès Guillot. Path planning versus cue responding: a bio-inspired model of

switching between navigation strategies. *Biological cybernetics*, 103(4):299–317, 2010.
60. Raja Chatila, Erwan Renaudo, Mihai Andries, Omar Chavez-Garia, Pierre Luce-Vayrac, Raphael Gottstein, Rachid Alami, Aurélie Clodic, Sandra Devin, Benoît Girard, and Mehdi Khamassi. Toward self-aware robots. *Frontiers in Robotic and AI*, 5(1):88–108, 2018.
61. Erwan Renaudo, Benoît Girard, Raja Chatila, and Mehdi Khamassi. Design of a control architecture for habit learning in robots. In *Biomimetic and Biohybrid Systems, LNAI Proceedings*, pages 249–260, 2014.
62. Erwan Renaudo, Benoît Girard, Raja Chatila, and Mehdi Khamassi. Which criteria for autonomously shifting between goal-directed and habitual behaviors in robots? In *5th International Conference on Development and Learning and on Epigenetic Robotics (ICDL-EPIROB)*, pages 254–260, Providence, RI, USA, 2015.
63. R. Dromnelle, B. Girard, E. Renaudo, R. Chatila, and M. Khamassi. Coping with the variability in humans reward during simulated human-robot interactions through the coordination of multiple learning strategies. In *Proceedings ofthe 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN 2020)*, Naples, Italy, 2020.
64. Adrien Jauffret, Nicolas Cuperlier, Philippe Gaussier, and Philippe Tarroux. From self-assessment to frustration, a small step toward autonomy in robotic navigation. *Frontiers in neurorobotics*, 7:16, 2013.
65. Giovanni Maffei, Diogo Santos-Pata, Encarni Marcos, Marti Sánchez-Fibla, and Paul FMJ Verschure. An embodied biologically constrained model of foraging: from classical and operant conditioning to adaptive real-world behavior in dac-x. *Neural Networks*, 72:88–108, 2015.
66. Diogo Santos-Pata, Riccardo Zucca, and Paul FMJ Verschure. Navigate the unknown: Implications of grid-cells 'mental travel' in vicarious trial and error. In *Conference on Biomimetic and Biohybrid Systems*, pages 251–262. Springer, 2016.
67. Simon Hangl, Vedran Dunjko, Hans J Briegel, and Justus Piater. Skill learning by autonomous robotic playing using active learning and creativity. *arXiv preprint arXiv:1706.08560*, 2017.
68. Martin Llofriu, Pablo Scleidorovich, Gonzalo Tejera, Marco Contreras, Tatiana Pelc, Jean-Marc Fellous, and Alfredo Weitzenfeld. A computational model for a multi-goal spatial navigation task inspired by rodent studies. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.
69. Dalia Marcela Rojas-Castro, Arnaud Revel, and Michel Menard. Rhizome architecture: An adaptive neurobehavioral control architecture for cognitive mobile robots' application in a vision-based indoor robot navigation context. *International Journal of Social Robotics*, pages 1–30, 2020.
70. Yevgen Chebotar, Karol Hausman, Marvin Zhang, Gaurav Sukhatme, Stefan Schaal, and Sergey Levine. Combining model-based and model-free updates for trajectory-centric reinforcement learning. *arXiv preprint arXiv:1703.03078*, 2017.
71. Anusha Nagabandi, Gregory Kahn, Ronald S Fearing, and Sergey Levine. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7559–7566. IEEE, 2018.
72. Muhammad Burhan Hafez, Cornelius Weber, Matthias Kerzel, and Stefan Wermter. Curious meta-controller: Adaptive alternation between model-based and model-free control in deep reinforcement learning. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.