Metalearning under uncertainty

Sami Beaumont

THÈSE DE DOCTORAT
Sorbonne Université

Spécialité : Sciences Cognitives

École doctorale nº158: Cerveau, Cognition, Comportement

réalisée

à l'Institut du Cerveau

sous la direction de Philippe DOMENECH et Mehdi KHAMASSI

présentée par

Sami BEAUMONT

Sujet de la thèse :

Méta-apprentissage en situation d'incertitude
Etude comportementale et computationnelle de l'adaptabilité chez l'homme

soutenue le 15 décembre 2022

devant le jury composé de :

M.	Summerfield Christopher,	Université d'Oxford,	Rapporteur
M.	Procyck Emmanuel,	SBRI, Lyon	Rapporteur
M.	Wyart Valentin,	Ecole Normale Supérieure, Paris	Examinateur
Mme	Scholl Jacqueline,	CRN, Lyon	Examinatrice
M.	Domenech Philippe,	ICM, Paris	Directeur de thèse
M.	Khamassi Mehdi,	ISIR, Paris	Directeur de thèse

Remerciements

Il me faut en premier lieu remercier les membres du jury, pour l'attention qu'ils ont portée à mon travail, leurs retours, critiques, et suggestions. Je tiens également à saluer le Professeur Renaud Jardri pour ses conseils et encouragements pendant ces 4 années lors des comités de suivi de thèse, aux côtés de Valentin Wyart.

Ce travail n'aurait bien sûr pas pu voir le jour sans la confiance et le soutien de Philippe Domenech et Mehdi Khamassi. Plus qu'une supervision, je garderai le souvenir de conversations riches, et d'échanges à la fois rigoureux et bienveillants. Je dois à Philippe de m'avoir initié au domaine, et accompagné pendant tant d'années jusqu'à la réalisation de cette thèse. Je lui suis reconnaissant de m'avoir confié ce projet, et d'avoir su me laisser libre tout en m'apportant un support constant.

J'ai eu la chance de bénéficier d'un cadre de travail formidable ces dernières années, au sein de l'équipe d'Eric Burguière à l'ICM. J'espère emporter avec moi quelques éléments de neurophysiologie, acquis dans une atmosphère chalereuse et ouverte tant humainement que scientifiquement. Merci particulièrement à Karim N'Diaye pour son travail à la plateforme PRISME, qui m'a permis de mener à bien ce travail au travers des pandémies et des pannes de réseau.

Les années passées à l'ICM me laisseront le souvenir de joyeux moments entre étudiants, avec ceux de l'équipe MBB — Nicolas, Jules, Emma, Chen, Tony et les autres — et bien sûr avec les infatigables Nerbies — Lizbeth, Pauline, Youenn, Hugues, Sarah, Eliott. Egalement Eliana et Oriana, aux côtés de qui j'ai eu le plaisir de partager cette expérience unique de la thèse, des demandes de financement, des deadlines, des bugs et de la dernière saison de GoT. Merci à Marine, pour son attention, son soutien et sa générosité. Merci à Lindsay, pour m'avoir guidé et encouragé, jusqu'au bout.

Je tiens à remercier toute l'équipe organisatrice de la BAMB !, pour leur incroyable summer school, dans un cadre si exceptionnel.

Sur un plan plus personnel, mes pensées vont vers mes proches, mes amis et ma famille qui ont parfois dû me partager avec cette abstraction qu'est le travail de thèse. Merci à Sinéad pour sa patience, son écoute et ses encouragements au quotidien. Merci à Colin pour m'avoir (bien malgré lui) donné autant d'inspiration. Merci à ma mère pour son aide indéfectible qui a rendu ce travail possible.

Table des matières

General introduction

> Perhaps to do without theories altogether is a *tour de force* that is too much to expect as a general practice. Theories are fun.
>
> B.F. Skinner, *Are theories of learning necessary ?* 1950

How do we learn to learn ? This question is at the core of our abilities to adapt flexibly in natural environments. On one hand, robust and reproducible mechanisms are needed to ensure efficient adaptation in an uncertain and volatile environment. On the other hand, the latent structure of the environment can vary over an unlimited range in number of dimensions, temporal dynamics or causal entanglements.

One can picture a PhD student trying to grasp some important questions in her field. She will formalize the problem, design an experiment and try to make sense of her results. At first, she will obtain unexpected outcomes and bugs, and by trial and error, with the help of many pilot experiments, she will finally get a reasonable and consistent result. This incremental process can apply to many real-life situations. It is studied in the laboratory in many different species. However, to be really adaptive, learning requires fine-tuning for each specific problem. Going back to the student, one can wonder : how confident should she be that her experimental design is not flawed in face of unexpected results ? Should her analysis be more exploratory, or mostly focus on her main hypothesis ?

The answers to those questions belong to the topic of **meta-learning**. meta-learning can involve the same mechanisms as learning itself : a PhD supervisor can incrementally learn how to improve research practices from the accumulated experience of all her students. But meta-learning might also require a qualitatively different approach : the supervisor could realize that, in her field, some questions need a specific type of experiment and analyses that are not ideal for other questions. The former illustrates the gradual and recursive nature of meta-learning. It assumes that the agent receives a continuous stream of data, and uses similar mechanisms to learn and learn to learn. The latter, however, portrays meta-learning as an active inferential process, leading to abrupt shifts as the agent uncovers the hidden cause(s) of the observed data.

Hence the dual face of meta-learning : it sometimes refers to continuous processes, and sometimes discrete inferences. It can be slow and gradual, but also a sudden revelation. It covers recursive low-level mechanisms but also high-level abstractions of the data generative process. The goal of this thesis is to improve our understanding of meta-learning, and attempt to reduce this apparent dichotomy. Importantly, while early experimental work investigated learning as a marker of animal intelligence in general, the 20th century saw the formalization of learning as a specific and low-level process that builds up links between elementary states, actions and outcomes. This view brought considerable progress in understanding animal behaviour and brain function, but partially disconnected learning theories from the study of other cognitive abilities, such as propositional reasoning, memory or categorization.

The following sections are devoted an historic of these theories and will review the state of the art of current models of associative (meta)learning (section 1). Then, I will present data and theories showing the involvement of discrete and high-level inferences that challenge the incremental and associative view (section 2). Interestingly, both bodies of work concern similar empirical phenomena, and involve overlapping brain regions, in particular in the prefrontal cortex. I will finally introduce our hypotheses and experimental strategy to give an unified account of existing evidence (section 3).

1 Associative learning and continuous adaptation

> Learning by having ideas is really one of the rare and isolated events in nature
>
> E. Thorndike, *Animal intelligence*, 1911

1.1 Theories of associative learning

During the 20th century, theories of animal learning evolved substantially. In fact, we shall see that the formalization of learning encompasses three distinct functions. The initial approach, framing learning as an associative mechanism between stimuli and actions, brought out the key notion of **reinforcement** [1, 2]. Then, accompanied by the birth of artificial intelligence, learning processes were thought of as solutions to the **credit-assignment** problem [3]. Finally, with the combined developments of neurophysiological measures and the Bayesian theoretical framework, learning became inseparable from the notion of **prediction error** [4].

Associative learning and conditioning

During the late 19th century, the American psychologist Edward Thorndike inaugurated the field of comparative psychology with a series of original studies on learning [5]. Using a carefully designed experimental apparatus, the puzzle-box, he tested whether

animals could learn to escape by having insight. He showed in several species that learning was a gradual process, and not an epiphany towards the solution, and that animals rarely generalized previous solutions to new problems. He thus came to the conclusion that "*Learning by having ideas is really one of the rare and isolated events in nature*" [5] (p. 284). His work had a significant influence for the nascent field of behavioral psychology, in particular what he coined the *law of effect* : learning is promoted through the outcomes of one's past actions. It is worth noting that Thorndike conceived elementary actions as reactions to stimuli (or "impulses"), and thus learning as the formation of associations between observable states and impulses, not ideas (pp.99-100). For him, animals could build a limited stock of associations via neurophysiological connections, that were gradually strengthened and selected via the repetition of favorable outcomes. He was very critical of earlier theories that compared animal learning to human learning, as he believed that the associations of ideas in humans were of a completely different nature to the associations between stimuli and impulses that he studied in other animals. (pp 125-126). In this respect, he fell in a teleological fallacy, by seeing ideas and rational inference as end results of evolution, and concluded, in a fit a false modesty : "*Amongst the minds of animals that of man leads, as a demigod from another planet, but as a king of the same race.*" (p. 294). This view of phylogeny as a hierarchy was common among early 20th century scientists, and Thorndike was, like many, a strong supporter of eugenics and gender inequality [6].

Thorndike's observations, though influential, raised several criticisms from behavioral psychologists. In particular, his statements that animals learned the solution of the puzzle box by incrementally refining actions sequences through positive or negative outcomes (which he called "satisfiers" and "annoyers") were not compatible with behaviorism, the dominant doctrine of the early 20th century psychology. Behaviorists were reluctant to give scientific consideration to private mental content, and gave more credit to another of Thorndike's law, the *law of repetition*. Unlike the law of effect, the law of repetition stated that associations could be strengthened by the simple recurrence of a given action, regardless of its results. Since the experiments were repeated in the same conditions, and always required the same sequence of actions to end, associations could form by statistical repetitions rather than subjective judgments. Margaret Washburn, a significant figure of comparative psychology, noted in her in 1908 textbook *The Animal Mind* [7] the lack of parsimony of behaviorists' explanation of learning, by naming one in particular, John Watson : "*Watson lays especial stress on the fact that the successful movements in puzzle-box and maze experiments have the advantage of frequency of performance. The successful movements are always performed, in every maze experiment, simply because the experiment continues until they are performed ; there is no such necessity that any particular unsuccessful movement should be performed in every experiment. Thus the successful movements, Watson thinks, owe their survival to the law of repetition. [...] This can only be done by ignoring such cases of learning as those where the frog ceased in one or two trials to snap at food when the snapping led to harmful consequences, or where the spider learned not to disturb itself at the sound of a tuning fork.*" (pp 276-277).

Another major criticism against Thorndike's law of effect was its retroactive nature.

Indeed, since an action could simply be generated through the stimulation of a neurophysiological connection, the retroactive modification of this connection via an external signal was a mystery. For example, the American psychologist R.H. Waters wondered : "*how can an effect or a consequence of an act have any influence, detrimental or beneficial, on the retention and repetition of the act which preceded and produced it ?*" [8]. While Thorndike's main hypothesis for associations formation was *selection* (from a predetermined and hardwired pool of associations), Waters suggested that learning occurs by the successive *production* of original behavior from some latent "disposition" that is progressively oriented towards the correct solution. Thus, despite several rejections from behaviorists, two major aspects of learning became undeniable : the abstract nature of the link between stimuli and actions, Waters's "disposition", and the subjective valence of action's outcomes, depending on the animal's goals and preferences, which alter this disposition.

During the first half of the 20th century, theories of learning progressively became more familiar to contemporary views. The concept of **reinforcement** is formally introduced by Clark Hull [1], as a more general term than Throndike's "effect". Hull noticed that a reinforcer is not always a consequence of the action, and that temporal contiguity is an important factor per se. He defined a *law of primary reinforcement* : "*Whenever a reaction (R) takes place in temporal contiguity with an afferent receptor impulse (s) resulting from the impact upon a receptor of a stimulus energy (S), and this conjunction is followed closely by the diminution in a need [...], there will result an increment,* $\Delta(s \longrightarrow R)$*, in the tendency for that stimulus on subsequent occasions to evoke that reaction.*" [1] (p. 71). Moreover, Hull explicitly framed conditioning as a special case of reinforcement learning. Indeed, more than puzzles and mazes, conditioning was the most influential paradigm to study learning within the behaviorist framework. One usually distinguish classical or Pavlovian conditioning from operant or instrumental conditioning. The former, initiated notably by Ivan P. Pavlov [9] describes the emergence of a physiological response (*e.g.*salivating) to a stimuli (*e.g.*a bell ring) when repeatedly associated with a reinforcement (*e.g.*food), while the latter focuses on the acquisition of conditioned actions (*e.g.*pressing a lever) [10].

In this context, another key concept due to Hull is nothing less than the delta rule, which he formalized mathematically, though not under this name, 30 years before the work of Rescorla and Wagner. In Hull's terms, the link between stimuli and actions is a "habit strength", that grows as a function of the number of reinforcers. He defines 3 factors for the growth of a "habit" ${}_sH_r$ [1] (p.114) :

1. a physiological limit (M)
2. the number of reinforcers (N) producing increments in habit strength $\Delta_s H_r$
3. a constant transfer factor (F) of the increment to the habit strength. In contemporary terms, a *learning rate*.

Contrary to Hull's habit strength, Robert Bush and Frederik Mosteller [11] proposed that learning was based on updating probabilities of actions after specific events. Interestingly, their formalization offered an explicit connection between learning processes

and *Markov chains*, *i.e.* the probabilities of actions or states at trial t only depends on the probabilities at trial $t-1$ and the event at t (the **Markov property**).

Finally, building on previous experimental and theoretical work, Robert Rescorla and Allan Wagner presented a model of Pavlovian conditioning, still in use today, generally referred to as the **delta rule** [12]. Their goal was to give a general theory accounting for several observations regarding associative learning with compound stimuli :

1. Blocking : When trained with a compound conditioned stimulus AX (*e.g.*a sound and a light signal), animals usually display conditioned response when tested for X alone. This response can be attenuated when A alone is also paired with reinforcement in parallel.
2. Conditioned inhibition : After an initial phase of associating the conditioned stimuli A and X with different reinforcement rates, extinction of the compound stimuli AX (*i.e.* its presentation with no reinforcement) has opposite effects when A was more strongly associated with reinforcements than when it was less strongly associated. When A was initially associated with a higher reinforcement rate, it will *inhibit* the reaquisition of X as a conditioned stimulus.
3. Magnitude effect : this inhibitory effect can be countered by increasing the magnitude of the reinforcement (*e.g.* the intensity of a shock).

Resorla and Wagner's proposal was not only based on their own experimental work, but also strongly influenced by Leon Kamin's *surprisingness* hypothesis [13, 14]. Kamin reported the blocking effect, and suggested that learning was driven by surprise : when previously conditioned with stimuli AX and A, there is no surprise left to learn the association between the reinforcement and X alone. Recent accounts on the role of surprise for learning and meta-learning will be discussed further, in sections 1.1 and 1.2. Rescorla and Wagner went back to Hull's formalism of associative strength, rather direct action probability as Bush and Mosteller did. Indeed, the existence of several interference effects between conditioned stimuli and the importance of the magnitude of the reinforcement, lead them to give a central role to the *total associative strength* V_{AX} of the compound stimulus AX. Just as Hull, they proposed that learning requires 2 parameters : a **learning rate** α and the subjective value of the reinforcement (equivalent to the asymptotic value) r. However, the local updates of associative strengths crucially depend on the total associative strength :

$$\Delta V_A = \alpha_A * (r - V_{AX}) \tag{1.1}$$

$$\Delta V_X = \alpha_X * (r - V_{AX}) \tag{1.2}$$

$$V_{AX} = V_A + V_X \tag{1.3}$$

Rescorla and Wagner's theory thus established two important concepts : 1/ learning requires some sort of representation of an abstract associative link and 2/ this link fundamentally depends on global variables. The total associative strength, as well as Kamin's *surprisingness*, were *ad hoc* proposals to account for the lack of learning when reinforcements were already associated with other conditioned stimuli. But it is also connected with the theoretical question of *credit assignment* which was already studied

by computer scientists at that time and led to major contributions for the reinforcement learning paradigm.

Artificial intelligence and the credit assignment problem

The study of conditioning with compound stimuli demonstrates the importance of identifying the source of reinforcement for learning efficiently. When stimuli are multi-dimensional or when rewards are delayed, it can be difficult to appropriately attribute credit to specific states or actions. This question has occupied experimental psychology from its inception. For example, Hull discussed the issue of distant rewards and secondary reinforcements [1], building on the work of Thorndike and Washburn, among many other early 20th century researchers. He proposed a mathematical model, similar to his law of primary reinforcement, to account for the exponential decay of the reinforcement strength with time.

However, from the 1950s and with the beginnings of artificial intelligence, normative approaches emerged. Marvin Minsky is probably the first author to explicitly frame learning as a credit assignment problem [3]. While he recognized the value of psychological models, he expressed doubt regarding their potential importance for the development of artificial intelligence and called for normative models. The most influential contribution at that time came from the field of optimal control and is due to the mathematician Richard Bellman [15]. Bellman coined the non specific term *dynamic programming* to designate his research program, in order to avoid mathematical jargon and attract resources from the military [16]. Dynamic programming generally refers to solving an optimization problem recursively, by breaking it into easier sub-problems.

Here the function to be optimized, called an *objective function*, corresponds to the expected value of a sequence of returns, or rewards, $\mathbf{R}$ when an agent applies the policy Π from the initial state s_0. At each time step t, the agent selects the action $a = \Pi(s_t)$ given the current state s_t, receives a feedback $r = R(s_t, a)$ and observes a new state $s_{t+1} = T(s_t, a)$. Π and T are the policy and the transition function respectively. The objective function from state s_0 writes :

$$V(s_0) = \sum_{t=0}^{T} \gamma^t R(s_t, \Pi(s_t)) \tag{1.4}$$

With $\gamma \in [0, 1]$ a **discount factor** that represents the agent's preference for close rewards over distant ones. Since each decision only depends on the current state and not the whole the history of past decisions, this process is a *Markov decision process*. This is a necessary assumption for the problem to be solvable recursively. Thus :

$$V(s_0) = R(s_0, a_0) + \sum_{t=1}^{T} \gamma^t R(s_t, \Pi(s_t)) \tag{1.5}$$

$$V(s_0) = R(s_0, a_0) + V(s_1) \tag{1.6}$$

The optimal policy solves the Bellman equation :

$$V^*(s) = \max_a \{R(s, a) + V^*(T(s, a))\} \tag{1.7}$$

Importantly, the solution to the Bellman equation is not a function or a specific value, but a policy. There exists no analytic or closed-form formula for the solution, hence subsequent developments of reinforcement learning can be viewed as algorithms aiming to solve the Bellman equation.

One of the first algorithms of this kind was proposed by Ian Witten [17]. The model specifies an estimate $W(s)$ of the expected discounted reward from state s with the following updating rule after a new reward r and the transition to the new state s' :

$$W_{t+1}(s) = W_t(s) + \alpha(r + \gamma W_t(s') - W_t(s)) \tag{1.8}$$

With $\alpha \in [0, 1]$ a **learning rate**. The agent then constructs the optimal policy by selecting the action that maximizes the expected return of the future state : $\Pi(s) = \arg\max_a\{W(T(a, s))\}$. One can recognize the delta rule augmented with the discounted expected return from the new state $\gamma W_t(s')$. This formalism will prove to be key in understanding the neurobiological mechanisms underlying associative learning (see section 1.1). Equation 1.8 belongs to a general class of reinforcement learning models called **temporal difference reinforcement learning** or *TD-RL*, a term coined by Richard Sutton [18]. Sutton, under the supervision of Andrew Barto, developed generalizations of the algorithm and proved its convergence. Their contribution was at the confluence of neuropsychology and computer science. They were influenced, for example by Harry Klopf's heterostatic theory of the brain [19] which states that neurons seek to maximize a hedonic state and not to stay in an equilibrium (homeostasis). Interestingly, several authors currently support the opposite view, as we will see later, by insisting on the notion of **surprise** rather than reinforcement. Sutton distinguished TD-RL models from other learning algorithms (*e.g.*backpropagation in neural networks) by the way they solve the credit assignment problem : "*The purpose of both backpropagation and TD methods is accurate credit assignment. Backpropagation decides which part(s) of a network to change so as to influence the network's output and thus to reduce its overall error, whereas TD methods decide how each output of a temporal sequence of outputs should be changed. Backpropagation addresses a structural credit-assignment issue whereas TD methods address a temporal credit-assignment issue.*" [18]

Among TD models, Christopher Watkins and Peter Dayan introduced Q-learning as a method to incrementally learn the policy, instead of deriving it from the expected values of states [20, 21]. Indeed, Q-learning allows optimization of the agent's policy locally, *i.e.* for each state-action association, while other TD methods require to recompute the value of every state in order to update the policy. The expected return of each state-action association is called a Q-value and is updated by :

$$Q_{t+1}(a, s) = Q_t(a, s) + \alpha(r + \gamma V_t(s') - Q_t(a, s)) \tag{1.9}$$

Where $V_t(s') = \max_a\{Q(a, s')\}$. While optimal action selection under TD models requires to maximise the expected return, Q-values are classically softmaximized, *i.e.* the

probability $\pi(a, s)$ of choosing action a in state s follows a Boltzman distribution [2] :

$$\pi(a, s) = \frac{e^{\beta Q(a,s)}}{\sum_i e^{\beta Q(i,s)}} \tag{1.10}$$

With $\beta \geq 0$ the **inverse temperature**. The introduction of this function in psychology is due to the mathematician Duncan Luce in his 1959 book *Individual Choice Behavior* [22]. Luce's proposition, known as *Luce's choice axiom*, aimed to capture the observation that the bias towards an option depends on the set of available options. This is mathematically translated by the normalization of some subjective quantity, the response strength $\mathbf{H}$, over the set of possible choices $\mathbf{A}$: $Pr\{a\} = \frac{H(a)}{\sum_{i \in \mathbf{A}} H(i)}$. The definition of the response strength as $H(a) = e^{u_a}$ makes Luce's model a multinomial logit model using the economic utility u [23]. In learning models, the utility of action a in state s is defined as a linear transformation of the Q-value $u(a, s) = \beta Q(a, s)$ (the constant term being omitted since it has no effect after the normalization).

Thus, the modeling of associative learning has crystallized around 3 free parameters : a learning rate, a discount factor and an inverse temperature. In further sections, we will review meta-learning models that dynamically adapt those parameters. However, it is worth noting that associative learning thus formalized is severely constrained. First, even though it can solve a temporal credit assignment problem, it does not explicitly represent the latent structure of the environment. Associative learning is necessarily local and limited to address more intricate problems. Second, this locality impose computational restrictions on its scalability to high dimensional problems. Finally, the very notions of action and state do not have unambiguous definitions, and other processes are necessary to specify the elementary basis of associative learning. This is not to say that reinforcement learning as a whole suffers from those issues. Many extensions and alternatives to TD-RL exist in order to address these limits. But the associative assumption, however fruitful, could also hinder our comprehension of learning.

Nevertheless, a central quantity in reinforcement learning models emerged : the **prediction error** : $\delta = r + \gamma V(s) - Q(s, a)$. It is ubiquitous to every associative models reviewed so far, from Hull's model to Q-learning, as it directly quantifies the update of behavioral tendencies as a function of perceived outcomes. The prediction error will prove essential for understanding the neural implementation of associative learning as well as for advancing computational models of learning.

Prediction and dopamine

The implication of the neurotransmitter dopamine in reinforcement learning is progressively understood during the 1990s, in particular with the work of Wolfram Schultz and colleagues through electrophysiological recordings in Macaque monkeys' midbrain. They were initially interested in the role of midbrain dopaminergic neurons in movement initiation, as previous studies showed inconsistent evidence of mibrain dopaminergic activity in monkeys during arm movement [24]. Moreover, other reports mentioned dopaminergic response to various stimuli whether they triggered movements or not. Thus,

Romo & Schultz [24] designed an experiment to answer two questions : 1/ how dopamine is implicated in movement initiation, and 2/ whether dopaminergic response to external stimuli reveals a perceptual function or is part of a motor reaction. The animals had to self initiate movement towards a box containing food or a wire with food at its end. Interestingly, phasic midbrain dopaminergic bursts were associated with motion only when food was present. When the animals reached an empty box or a bare wire, no dopaminergic bursts were recorded, and dopaminergic activity was even sometimes suppressed. Subsequent studies confirmed and extended the understanding of the role of dopamine in reward prediction using learning and conditioning paradigms [25, 26], until phasic dopaminergic bursts were eventually interpreted as reward prediction errors signals, as formalized by TD-RL models [4].

The functions of the projection areas of midbrain dopaminergic neurons rapidly came into scrutiny. An early hypothesis suggested that domamingergic phasic activity serves as a general teaching signal by changing synaptic connectivity in two major structures : the ventral striatum and the orbitofrontal cortex [27]. The former would encode reward expectation, that would be modulated by incoming reward prediction errors, while the latter would discriminate the qualitative origin of undifferentiated reward related signals.

Ventral striatum and actor-critic models

The role of the striatum for value-based action selection is known since the 1960-70s, as a bridge between so called *"limbic"* and *"motor"* systems in the mammal brain [28]. In particular, dopaminergic projections from the ventral tegmental area (VTA) to the ventral part of the striatum were soon considered to be essential for the initiation of a motivated action [28, 29]. With the rise of TD-RL models, the ventral striatum (VS) was thought to play the role of an *adaptive critic* [30, 31].

Actor-critic models are a general class of biologically plausible models where adaptive behavior emerges through the interplay of two distinct systems : a *critic* that learns to anticipate rewards, and an *actor* that uses this information to trigger appropriate behavioral responses [30]. Several models of the basal ganglia place the VTA to VS connections as the critic, *i.e.* the locus of reinforcement learning *per se* [30, 31, 32, 33]. Indeed, the time-course of dopaminergic release in the ventral striatum has been found to correlate with reward prediction error in rodents [34, 35], as well as the ventral striatum's BOLD signal in functional MRI (fMRI) in humans [36, 32, 37, 38]. However, playing the role of the critic implies that the ventral striatum does not merely encode reward prediction errors, but also expected rewards signals. Evidence regarding anticipatory reward signals in the ventral striatum is less clear : the timing of such anticipatory signal is less temporally accurate in rodents than it is predicted by TD-RL models [39]. Interestingly, extending the model to assume a distribution of critics instead of a single critic component, *i.e.* implementing uncertainty about the true state of the world, allows to solve this discrepancy [39, 29]. Section 2.2 will address such models in details.

Positive and negative valence

While VTA to VS phasic dopaminergic signals are consistently associated with positive reward prediction errors and positive expected values [40, 37, 41], the brain regions involved in negative reinforcement learning are still under debate. Early hypotheses

suggested the implication of another neurotransmitter, serotonin, that would act as an opponent system to dopamine [42]. However, empirical evidence is more consistent with a alternative computational interpretation for serotonin's function, as we will discuss later [43]. Other studies point towards several cortical areas for negative prediction errors and negative expected values, in particular the amygdala [41] and the anterior insula [37, 44, 45], though the latter might encode unsigned prediction errors, *i.e.* salience [46] or risk [47, 48]. Consistent electrophysiological evidence in non human primates also direct towards the habenula as a specific input region to the VTA for negative prediction errors [49, 50, 51, 52].

Nevertheless, it is worth noting that associative learning *per se* does not imply such a strong dichotomy between positive and negative domains. Hence, this biological distinction could suggest alternative computational interpretations. Indeed, negative prediction errors correspond to unexpected and undesired outcomes which could require to more radical adaptations than their positive counterpart. Models that explore the idea of detecting changes and leaving ineffective strategies will be detailed later, in section 2.

Orbitofrontal and ventromedial prefrontal cortex

As mentioned earlier, the orbitofrontal cortex (OFC) was initially thought of as a downstream structure that would relate the generic dopaminergic signal to specific goals [27]. In an fMRI paradigm aiming at disambiguating goal value, decision value and prediction errors, [38] showed that activity in the medial OFC correlated with goal values, while the central OFC and the ventral striatum were associated with decision value and prediction errors, respectively. Consistent with this idea, overlapping patterns of activity in the ventromedial prefrontal cortex (vmPFC) were identified for different types of rewards, suggesting the existence of a common neural currency for value-based decision making [53, 54, 55]. Moreover, lesion studies confirmed the causal implication of this region for value-based decision making in humans [56] and non human primates [57].

However, the OFC's function could go beyond encoding previously learned values. In a behavioral task elegantly discriminating behavioral responses using previously learning (cached) values from dynamically inferred values in rats, lesioning the OFC only hurts the latter [58]. This motivated the hypothesis of a more abstract function of the OFC, putatively as a cognitive map of the task space [59, 60] or a representation of the current policy's reliability [61]. Such models will be discussed at length in section 2.2.

So far, we presented the origins, the success and the limitations of associative learning models for understanding animal behavior. The global picture is strongly based on the assumption of locality between elementary stimuli, actions and rewards, through iterative prediction errors likely implemented by the dopaminergic phasic activity from the VTA. However, in order to stay adaptive, such mechanisms require a fine-tuning to the context, the goals and the timescale of the problem at hand. In the next section, we present several lines of work that extend associative reinforcement learning to such meta-adaptations.

1.2 Reinforcement learning and meta-learning

Temporal difference reinforcement learning minimally requires a subjective representation of the associative value between a state and an action, and the valuation of a reinforcement in order to compute a prediction error. Around this minimal process, meta-learning can adjust the **timescale** (section 1.2), especially the learning rate, the **exploration biases** (section 1.2), and the **contextual modulation** of the valuation (section 1.2).

Learning rate and timescale adaptation

What is a learning rate and how is it biologically implemented ? One of the earliest interpretation of the learning rate is due to John Pearce and Geoffrey Hall [62, 63]. The **Pearce-Hall model** was developed in the context of conditioning as a successor to the Rescorla-Wagner model. While the latter focused on the strengthening of associative links via reinforcement itself, the Pearce-Hall model aimed to capture the variability from the conditioned stimulus. Indeed, reinforcement learning could depend on how much attentional resources were allocated to the CS, and subjects need to learn how much they should attend the CS in order to properly learn the associative link later. If the environment comprises N stimuli, the Pearce-Hall model defines the learning rate as [62] :

$$\alpha_t = |r_{t-1} - \sum_{i=1}^{i=N} V_{i,t-1}| \tag{1.11}$$

With r_t the reward obtained at trial t and $V_{i,t-1}$ the expected value associated with stimulus i at trial $t-1$. Note that in the case of a single stimulus presentation this corresponds to setting the learning rate as the unsigned prediction error of the previous trial. Equation 1.11 predicts that in stable environment (*i.e.* when the reward distribution does not vary), the learning rate inevitably reaches O, and thus accounts for experimental effects such as blocking and conditioned inhibition. Another variant of the Pearce-Hall model allows for smoother adjustments of the learning rate [63, 64] :

$$\alpha_t = (1-\eta)\alpha_{t-1} + \eta|r_{t-1} - \sum_{i=1}^{i=N} V_{i,t-1}| \tag{1.12}$$

With $\eta \in [0,1]$ a meta-learning rate. In this variant the learning rate varies as a low-pass filter of the unsigned prediction errors. Similar ideas have been proposed in the context of non-stationary reinforcement learning problems [65, 66].

Several neural implementations of the learning rate have been proposed since the 2000s. One of the earliest suggestion is supported by the effects of the neurotransmitter acetylcholine in memory and synaptic plasticity in the prefrontal cortex and the striatum [67]. Early recordings in monkeys' striatum revealed an increasing response of tonically active neurons during the course of conditioning, compatible with the recruitment of cholinergic interneurons that may modulate the effect of dopaminergic projections [68, 69]. As we will see later, an alternative computational function of acetylcholine has

been proposed in a Bayesian framework : the signaling of *expected uncertainty*, *i.e.* the expected mismatch between predictions and actual outcomes [70, 71]. This is analogous to defining the learning rate as the moving average of the unsigned prediction errors in equation 1.12.

In parallel, serotonin has been suggested as a candidate for a biological implementation of the learning rate, though its role in meta-learning is still in debate. Doya [67] proposed that serotonin represents the discount factor, although there is still a lack of evidence to support such a hypothesis [72]. Alternatively, pharmacological manipulation through serotonin reuptake inhibitors in humans revealed the involvement of this neurotransmitter for enhancing learning, independently of outcome valence [43]. Moreover polymorphisms of the serotonin transporter genes are associated with shifting behavior after a loss [73], which might partly be interpreted as a learning rate.

The activity of another neurotransmitter, norepinephrine, has been associated with learning rate modulation via unsigned prediction errors [64, 74, 75]. For example, Jepma and colleagues [76] showed that pharmacological mitigation of norepinephrine affected the learning rate during a probabilistic learning task in humans. More specifically, the noradrenergic projections from the locus coeruleus (LC) to the dorsal anterior cingulate cortex (dACC) are regularly associated with latent updates of internal models [64, 74, 77, 78, 79, 80]. As I will discuss in detail in sections 1.3 and 2.1, the dACC is a serious candidate for uncertainty-related learning rate variations [81, 82]. In this regard, activity in the dACC has been associated with the expected prediction error, given the his [83].

Finally, other studies suggest network-wise implementations of the learning rate, instead of a one-to-one mapping with a neurotransmitter. For example, reward memory traces in monkeys' PFC reveal a reservoir of time constants [84]. This is equivalent to a distribution of learning rates, that has been shown to drastically improve TD-RL algorithms [85]. Moreover, a variable learning rate could emerge from the adaptability of synaptic plasticity — or *metaplasticity* [86]. Biological evidence for metaplastic synapses remain indirect, though it might be implemented in cortical and subcortical areas, via several neurotransmitters such as norepinephrine, acetylcholine and dopamine [86].

Thus, the variety of putative mechanisms for controlling the learning rate could be explained by complex interactions between multiple neural systems across cortico-subcortical networks. It could also suggest that the linear scaling of updates as predicted by TD-RL models, is inherently limited to account for learning in general. In particular, we will discuss later (section 2) how the LC to dACC noradrenergic projections might convey abrupt reset signals instead of continuous adaptations [77, 80].

Random and directed exploration

In the reinforcement learning framework, exploration refers to any behavior not aiming at maximizing expected returns. In large environments, exploration is necessary to gain information and ultimately exploit the best options leading to an **exploration-exploitation dilemma** [87]. One classically distinguishes random (or undirected) from directed exploration [88]. The former comprises sampling schemes where suboptimal behavior arises from noise, *e.g.*softmax (Boltzman distribution) or ϵ-greedy (semi-uniform

distribution) action selection [2], whereas the latter involves the explicit computation of an information bonus on less chosen options. For example the upper-confidence-bound (UCB) scheme gives an additional value to each option, depending on the number of times they were chosen. Hence, the total utility of an action a at trial t is computed as [2] :

$$U(a)_t = Q(a)_t + c\sqrt{\frac{\ln t}{N_t(a)}} \tag{1.13}$$

Where $Q(a)_t$ and $N(a)_t$ are the Q-value and the number of previous occurrences of action a at trial t respectively, and c is a constant. Note that the reward-based and the information-based parts of the utility are independent from each other : this allows to discriminate their respective neural correlates. In a 3-armed bandit task, where subjects had to track the continuously fluctuating values of 3 options and choose the best, BOLD activity in the OFC and vmPFC correlated with reward based utility, while the frontopolar cortex and the intraparietal sulcus correlated with information-based utility [89].

There is convergent evidence that random and directed exploration co-exist and are contextually modulated [90, 91, 92]. Using a version of the bandit task with varying time horizons (*i.e.* the number of trials left), Wilson and colleagues [90] showed that human subjects are more exploratory, via random and directed exploration, with larger horizons. Modulation of directed exploration, via the tracking of relative uncertainty between options, has consistently been associated with the rostral part of the lateral PFC [93, 94]. Interestingly, this region is also involved in counterfactual inference about alternative options (Boorman et al 2011) or global strategies [95]. The neural implementation of random exploration, however, is probably more distributed [92]. The involvement of norepinephrine has been suggested repeatedly [67, 96, 97], but pharmacological manipulations via norepinephrine reuptake inhibitors did not confirm its causal implication [98, 99].

Alternatively, random behavior may arise not only from random action selection, but rather from noise in the update itself, *i.e.* noisy prediction errors [100, 101]. Indeed, suboptimal choices may be thought of as optimal under finite computational precision [102, 103]. Interestingly, the contribution of selection noise and imprecise inference to random exploratory behavior are separable, since selection noise only affects the current trial while inference noise is carried over from trial to trial. This allowed Findling and colleagues [100] to quantify the contribution of learning noise in a bandit task, revealing that it was the most important factor of behavioral stochasticity. Moreover, fMRI activity in the dACC correlated with learning noise, as well as pupillometry, an idirect marker of noradrenergic activity [100]. This original model also provided an alternative interpretation of mixed results from pharmacological manipulation of norepinephrine [101], giving strong arguments in favor of interpreting NE inputs on dACC as noisy learning signals.

However, noradrenergic projections to the dorsomedial prefrontal cortex are also associated with abrupt switches to random exploration in rodents [104] and primates [77]. Once again, this might suggest the extension of reinforcement learning models towards

non-linear global strategic switches, as we will discuss in section 2, but could be analogous to a variable inverse temperature of the softmax transformation. Indeed, transient decrease of this parameter would lead to a randomization of behavioral responses, as if previously learned values had been reset. In this regard, an active adjustment of the inverse temperature based on the average reward received so far has been proposed to solve the exploration-exploitation dilemma [105, 106, 107]. For example, the inverse tempoerature β could be adjusted depending on the difference between a short and a long range moving average of the rewards [107] :

$$\beta_{t+1} = \beta_t + \lambda(\bar{\bar{R}}_t - \bar{R}_t) \tag{1.14}$$

$$\bar{R}_{t+1} = \bar{R}_t + \eta(r_t - \bar{R}_t) \tag{1.15}$$

$$\bar{\bar{R}}_{t+1} = \bar{R}_t + \eta(\bar{R}_t - \bar{\bar{R}}_t) \tag{1.16}$$

Where $\bar{R}$ and $\bar{\bar{R}}$ are the first and second-order reward moving average respectively. Such a function might be implemented by the tonic dopaminergic activity in the striatum, that modulates action selection in the basal ganglia [108]. In support of this model, sustained dopamiergic activity has been found to correlate with reward uncertainty [109], and the modulation of dopaminergic activity in rats impacts random exploration [110].

Contextual adaptation

In addition to dynamically adjusting hyperparameters, efficient associative learning also requires context-specific adaptations. Here, I briefly review three common adaptations : learning from counterfactual outcomes, scaling the value function, and selecting relevant features of the stimuli.

Counterfactual learning

In many situations, available options are not independent from each other and the outcomes of a choice are informative to reevaluate non chosen actions. There is substantial behavioral and neuroimaging evidence that humans learn from such counterfactual signals. Fictive prediction errors, *i.e.* the difference between the actual outcome and the best possible outcome, are associated with BOLD activity in the ventral striatum [111]. They are distinct from counterfactual errors, *i.e.* the difference between the fictive outcome of non-chosen options and its expectation, that have been correlated with a broad prefrontal network encompassing the frontopolar cortex and the dorsomedial PFC [112, 113]. Furthermore, EEG data reveals that factual and counterfactual prediction errors follow distinct temporospatial pathways [114]. However, some behavioral data and computational analysis suggest an alternative view, that counterfactual learning emerges from confirmation bias rather than a separate learning system [115]. In a modified reversal learning task where subjects could observe outcomes for non-chosen options, computational modeling revealed a reversed valence-asymetric pattern for factual and counterfactual outcomes : subjects displayed higher learning rates for positive factual outcomes, whereas negative counterfactual outcomes were associated with higher learning rates. Interestingly, this pattern was explained more parsimoniously by confir-

mation bias, as positive factual outcomes and negative counterfactual outcomes have in common to confirm prior beliefs [115].

Range adaptation

Subjective values computed in a common neural currency [53, 54] are not strictly a function of objective values, but are scaled according to a context [116, 117]. Palminteri and colleagues [118] used a probabilistic learning task with 2 contexts, either with only positive (rewards) or negative (punishments) outcomes, and showed that the learning rate was similar in both contexts. However, post-learning tests where subjects are asked to choose from 2 options drawn from both contexts, revealed a non trivial effect : less punishing options were preferred over less rewarding options, despite their disadvantage in absolute value. This can be explained by range adaptation, *i.e.* the scaling of subjective values to the average value of the context. Indeed, mild punishment in the negative context are more appealing than low rewards in the positive context. These results extend to more contexts with variable magnitudes and probabilities of objective outcomes [119, 120], and suggest specific neural mechanisms for range adaptation, such as divisive normalization [121].

Nevertheless, relative valuation might also emerge from direct policy learning, as action selection only requires the identification of the most valuable option, independently from its absolute value [117]. Post-learning assessments using options from contexts with different reward magnitudes revealed that subject were biased towards options from the highest valued context [120, 117]. This highlights an interesting trade-off between global and local inferences. From the perspective of associative learning, range adaptation is a non trivial phenomenon that requires specific extension to be accounted for. In contrast, policy learning takes a global perspective from which range adaptation is a natural consequence, though arousal effects from higher local rewards might also bias inter-contextual choices. Other contextual effects support the importance choices valuation over policy learning in such tasks. For example, when required to select between two items while seeing a third, unavailable one, subjects tend to modify their choices depending on the value of this distractor [122]. Moreover, in a sequential value-based decision making task, subjects' choices were biased by the order of the options, suggesting that subjective values were influenced by repeated pairwise comparisons [123].

Selective attention and task-relevant information

Besides the attentional interpretation of the learning rate (see section 1.2), the valuation of available options is crucially dependant on the attentional frame [124, 125]. Moreover, attention interacts with learning by enabling the identification of structure in very rich environments [126, 127]. Indeed, naive reinforcement learning algorithms are severely slowed down when tracking multidimensional stimuli. This *curse of dimensionality* is an integral part of any credit assignment problem. Thus associative learning requires a parallel process to selectively attend relevant features of the environment [128]. Niv and colleagues [129] compared several candidate models for this process : Bayesian inference of relevant features, Bayesian selection of a random subset of features, and reinforcement learning of features relevance re-using the reward prediction error. They found evidence for the latter, suggesting that prediction error not only serves as an associative learning

signal, but can also shape more abstract representations [128]. Indeed, task-relevant information can be encoded as a by-product of learning itself, without necessarily relying on attentional effects [130]. Along these lines, authors have suggested that the subcortical dopaminergic system trains the prefrontal cortex to develop a task-specific learning algorithm, making the whole cortico-subcortical network a *meta-reinforcement learning* system [131].

1.3 Bayesian inference

In the early 2000s, developments of reinforcement learning around the notions of prediction, surprise and uncertainty, led to a conceptual re-framing of sequential learning as Bayesian inference [132]. Indeed, the idea that the brain needs to actively filter out uncertainty to extract relevant perceptions out of sensory data goes back to Hermann von Helmholtz's suggestions in the late 19th century. In the Bayesian framework, beliefs are formalized as conditional distributions over latent variables. Bayes theorem states that, after an observation, the posterior belief is proportional to the product of the prior belief and the likelihood of the observation :

$$P(Z|D) \propto P(Z) \times P(D|Z) \tag{1.17}$$

Assuming a stream of data D generated by some latent process Z, applying Bayes theorem leads to regression of the location, or mode, of the posterior distribution towards the true generative process Z^*, while its scale, or variance, represents the residual uncertainty due to estimated perceptual noise given the limited amount of data. The Bayesian brain hypothesis places Bayesian inference as a general computational framework to explain cognitive abilities and brain functions [133, 134]. Its application to learning initially informed automatic categorical learning in infants [135] and adults [136], and sensorimotor control [137] where agents have to dynamically adapt their movements to noisy feedbacks. In this section, I will review later developments of Bayesian theory for meta-learning in associative learning tasks.

Meta-learning and uncertainty

One specificity of the Bayesian framework is its focus on the precise dissection of uncertainty. The work of Yu & Dayan [70, 71] continued the search for a mapping between learning parameters and neuromodulators from the perspective of the multiplicity of uncertainty sources. They proposed the now classical functional distinction between **expected** and **unexpected** uncertainty, the former being the direct consequence of noisy and partial observations while the latter is caused by unpredictable changes in the environment. Based on experimental data, and particularly from attentional tasks, they proposed that acethylcholine plays a role in monitoring expected uncertainty, while norepinephrine is involved during task-shifting events, *i.e.* unexpected uncertainty [71]. Note that this is only partially consistent with hypotheses in the reinforcement learning framework : while acethylcholine has been associated with the learning rate (*i.e.* the weight of new observations compared to the memory of previous events), suggestions

regarding norepinephrine, and more generally noradrenergic projections from the LC to the dACC, are less consensual.

The dACC was first associated with error monitoring [138] and cognitive conflict [139], but later work revealed its critical implication for motivated behavior [140, 141] and reward prediction [142, 143, 144, 81]. In the Bayesian framework, a change in expected rewards corresponds to volatility, or unexpected uncertainty, and requires to modulate the learning rate in order to quickly discount irrelevant reward history. Using a Bayesian model of human behavior in a probabilistic reversal learning task, Behrens and colleagues showed that human learning is adjusted as predicted by the normative model, *i.e* the learning rate increased in volatile environments, and that dACC BOLD signal correlated with volatility monitoring [82]. However, the specificity of this signal to unexpected uncertainty remains unclear : activity in the dACC has been correlated more generally with unsigned reward prediction errors [74, 83] which conflate expected and unexpected uncertainty. Moreover, as we discussed earlier, LC to dACC projections might produce noise in the learning process itself [100], that could be misinterpreted as volatility [145].

In order to better understand the functions of the dACC and noradrenergic outputs from the LC, Silvetti and colleagues [75, 146] designed an alternative probabilistic reversal learning task including 3 conditions manipulating systematically volatility (rate of change) and noise (feedback variability). Remarkably, while noradrenergic activity inferred from pupillometry was specifically higher in the high volatility condition [75], fMRI activity in the dACC was associated with both feedback noise and volatility [146], suggesting a functional duality of the dACC in order to track multiple sources of uncertainty and act accordingly [78].

Hierarchical inference

In the Bayesian framework, a learning model is similar to a classifier : each observation is associated with a (latent) source, or generative process. This process allows the agent to infer which generative process is more likely and to adopt an optimal policy. Generative processes are classically structured hierarchically, from high-level causes to low-level observations. The **Kalman filter** is the most common instance of such models [147, 148]. The hierarchy is minimal, with only one level above the observable data $\mathbf{x}$, assumed to come from a Gaussian distribution : $\mathbf{x} \sim \mathcal{N}(\mu, \sigma^{\mathbf{2}})$. In the one dimensional case, after each observation x_t the estimate on the current latent state μ_t is updated as :

$$\mu_{t+1} = \mu_t + K_t(x_t - z_t) \tag{1.18}$$

Note that equation 1.18 corresponds to a delta-rule, with a learning rate K_t called a *Kalman gain* :

$$K_t = \frac{\nu_t^2}{\nu_t^2 + \sigma_{\text{obs}}^2} \tag{1.19}$$

$$\nu_{t+1}^2 = (1 - K_t)(\nu_t + \sigma_{\text{gen}}^2) \tag{1.20}$$

With ν_t^2 the variance of the current estimate of the generative process, σ_{obs}^2 the variance of the observation noise and σ_{gen}^2 the variance of the true generative process (*i.e.* the volatility). Both observation noise and volatility are assumed to be known, hence the learning rate corresponds to the optimal ratio between expected uncertainty and total uncertainty (expected and unexpected).

Nonetheless, the generative process can be far more complex, involving several layers of latent variables, each corresponding to the distribution over lower level variables. Such a network, connecting high-level latent causes and low-level observations, with possibly many intermediate nodes, is a *Bayesian network*. Without any constraints, exact inference on a Bayesian network is intractable [149]. Peter Dayan and colleagues proposed an approximate model, *the Helmholtz machine*, able to address inference over many latent hypotheses, using a computational trick : the *variational free energy* [150]. Though an extensive exploration of this principle is beyond the scope of the thesis, one can summarize the general idea as the transformation of a tedious inference problem into a optimization problem. As we discussed earlier, optimization problems can be solved with dynamic programming, which is both computationally accessible and biologically plausible.

This led to the formalization of (meta)learning as hierarchical Bayesian inference through, for example, a hierarchical Gaussian filter [151], or more recently a volatile Kalman filter [152]. Such models have several advantages. Coming from normative principles, they offer a natural explanation for evolving meta-adaptability. Indeed, hierarchical Bayesian filters will learn the latent structure of their environment by aiming at homeostasis. In other words, the model convergence is guaranteed by the reduction of surprise through the network. Hence, it will learn to predict upcoming observations as a by-product of inner stability [153]. This is completely opposite to earlier views, for example Harry Klopf's, of neurons as being *"hedonic"* and learning as a heterostatic process. It also dispenses with defining an explicit reward or reinforcement, in favor of the notion of surprise which applies more generally to any state that can be visited by the agent [154]. In addition, the generality of Bayesian networks, and the recursive nature of hierarchical inference makes heuristics and *ad hoc* explanations of meta-learning processes unnecessary. However, the variational approximation, and more generally the free energy principle are no panacea. Regarding computational constraints, such approximate methods cannot make unbounded Bayesian inference tractable [155]. The contribution of variational free energy lies in the reduced cost of the inference process and the factorization of the latent space, that make inference accessible by conventional optimization methods. Nevertheless, it still requires structural constraints over the latent space [155, 156]. Furthermore, neurobiological evidence supporting of variational inference in the brain is still lacking and some authors have argued that the free energy principle is more a general framework for building computational models than a refutable theory [157, 158].

Reinforcement learning and Bayesian inference

So far, I reviewed reinforcement learning and Bayesian inference separately. Though they are distinct historically and mathematically, several works have demonstrated their compatibility and convergence for studying the brain [154, 159, 160]. First, both theoretical frameworks aim at explaining different levels of Marr's classical tripartition [161] : Bayesian inference is mostly a computational level theory, prescribing normative solutions about how agents should solve specific problems, while reinforcement learning is mainly concerned with the algorithmic level of explanation, describing how agents actually solve them [1]. Second, they make similar predictions regarding the functional role of brain regions such as the dACC (*monitoring uncertainty* or *integrating unsigned prediction error* describe the same computational function). Finally, several authors have studied their mathematical convergence in specific conditions [160, 162]. For example, as illustrated with the Kalman filter, the assumption that observations are drawn from a Gaussian distribution simplifies Bayesian inference and allow linear updating of posterior beliefs similar to reinforcement learning models. In this case, one can use a Kalman filter as a basis model for meta-learning and add complementary mechanisms from the reinforcement learning literature, such as UCB or a dynamical inverse temperature [163]. Moreover, several authors have extended standard reinforcement learning models to handle distributions over values, rather than point estimates, effectively bridging the gap with Bayesian inference [85, 164, 165].

At the neurobiological level, these connections make fruitful predictions regarding the role of *e.g.* phasic dopaminergic activity. Recent work showed that, rather than prediction errors from point estimates of the exact state value as in classical TD-RL, dopamine might reflect the prediction error from the expected value of uncertain latent states [39, 29, 166]. For example, the correlation between the firing rate of dopaminergic neurons and the timing of rewards shows a different pattern depending on whether the reward distribution is deterministic or probabilistic [167]. Moreover, in an experiment where mice were trained in two contexts with different reward magnitudes, dopaminergic activity did not correlate monotonically with prediction errors, but showed a strong interaction between the reward size and the probabilistic representation of the underlying context [168].

Once again, this illustrates the proximity between learning and credit-assignment. Indeed, in natural environments, agents likely incorporate beliefs about the possible causes of their observations into their learning processes, since perceptual information is not necessarily as clear as experimental stimuli. Far from Thorndike's conclusions that inferences are absent from animal learning, the development of theories of meta-learning reveals that every step towards association formation requires inferences, from the tuning of hyperparameters to the very identification of the current (hidden) state. Going one step further, one can suggest that associative learning itself is a form of inference over latent states. In the next section, we will examine experimental results

1. This is mostly true for RL models in cognitive neuroscience, whereas most of contemporary RL models used in machine learning are derived as solutions of the Bellman equation for optimal control.

and computational models that challenge two aspects of associative learning as we have described it so far : 1/ its gradual nature and 2/ its local nature (by opposition to global inferences over abstract objects). Indeed, as we shall see, discrete adaptations and global inferences about the latent structure of the environment are essential ingredients of meta-learning mechanisms.

2 (Meta)learning as inference over latent states

2.1 The case against gradual associative learning

Despite the influence of the law of effect and later theoretical developments focused on the gradual nature of learning, the issue regarding discontinuities in the learning curve has been hotly debated throughout the 20th century. After discussing empirical evidence that gave rise to these debates (section 2.1, I will review meta-learning models built on **adaptive change-point detection** (section 2.1), and present the corresponding neurophysiological data (section 2.1).

The discontinuity of the learning curve

During the 70s and 80s a series of experiments, mostly performed with pigeons and rats, revealed what was later named *the matching law* [169]. The animals had to choose between two sources of rewards, with variable interval schedules. By pressing a lever or moving to another side of their box, they could switch from one schedule to the other. Intervals between reinforcements were independently sampled from Poisson distributions. Crucially, when a reward was scheduled in the absence of the subject, it was held and delivered as soon as the subject came back. Hence, when one side yields a better reinforcement rate than the other, the optimal strategy to maximize the total amount of reward is to stay on the best side and to sometimes check the other side to collect any pending reward. Interestingly, even when the differences between the two schedules were important, animals deviated from the optimal strategy by alternating between the two schedules — an empirical result called *the matching law* : the ratio of time spent in the two schedules almost perfectly matches the ratio of the two reward rates [169, 170]. One of the explanations given for this phenomenon was associative reinforcement learning : an animal would gradually learn the reward rate of each schedule and choose to press the lever proportionally. However, Gene Heyman proposed another theory : that the time spent in a schedule is an "unconditioned" effect, *i.e.*the decision to change schedule is independent of the amount of reward collected, and is only elicited by the frequency of reinforcement [170]. This idea was also put forward by John Gibbon [171, 172] who proposed that, instead of updating a belief on the value of both schedule, subjects store previously observed time intervals they could later sample to select the schedule with the shortest one. Crucially such a model assumes that the subject's behavior does not have any effect on the observations, since the memorized intervals are supposed to represent the actual underlying distributions. This is what Heyman means by "unconditioned" effect, *i.e.* independent from perceived consequences of one's action, and what Charles

Gallistel later called a feedforward model : "*The locus of reinforcement's effect is not in the mapping from perceived situations to actions nor in the mapping from actions to the amounts of reward they are expected to produce ; rather, it is in the subject's representation of income histories.*" [173].

This original model makes an important prediction, which is fulfilled according to Gallistel [173, 174] : since there is no feedback mechanism to update an associative link between stimuli and actions, the feedforward model can adapt quite rapidly (in fact as quickly as optimally possible) to any change in the variable interval schedules distributions. Gallistel and colleagues showed that rats can adapt in a couple of trials to a latent change in the probabilistic distributions, hence presenting the rat as "an almost ideal detector of changes" [173]. Later, they extended these results to the acquisition of conditioned behavior in the pigeon [175] : by using cumulative curves of actions (pecks) for individuals instead of averaging trial by trial behavior over the whole population, they show abrupt transitions in pigeons behavior. They conclude that the learning curve is an averaging artifact, and that learning is not a gradual process but an abrupt detection that some decision variable reached a threshold. Interestingly, they refer to information theory and link learning experiments with detection experiments [174, 175]. Indeed, the sequential integration of information is still required, but instead of a representation of associative strength, it enters an all or nothing type of decision. This is analogous to particle filtering with a small number of particles at the individual level, that will behave as a Bayesian filter at the ensemble level [176]. It also makes fruitful bridges between learning theory and categorization, especially in the case of one-shot learning, as discussed below in section 2.2.

Framing learning as change-point detection still has several limitations and drawbacks. First, while suggestive, the experimental data and analyses alone cannot formally refute gradual associative learning as long as subjective values are updated quickly enough and undergo a non linear transformation that can account for the fast transition to asymptotic performance. This issue also reflects a lack of direct comparisons between change-point models and continuous models in the literature [177], or worse, that they could not be distinguishable from behavioral observations collected in simple tasks [178]. Second, it remains unclear whether change-point models account for learning in general, or only in experimental designs that favors abrupt detection, such as variable intervals schedules or acquisition of conditioned behavior. Finally, even if change-point models are framed as feedforward models, they still require feedback adaptation to the time frame of the task (*i.e.* the memory depth of past events), which Gallistel acknowledged as an open question [173].

In the upcoming section, we review hierarchical Bayesian models of change-point detection that optimally adjust their timescale from available information. However, despite the significant insights they provide, they do not bring unambiguous evidence in favor of discontinuities in learning and meta-learning processes.

Hierarchical modeling of latent change-points detection

Without pre-specified constraints, latent change-point detection is a difficult problem for living organisms for two main reasons. First, while offline detection (*i.e.* after observing the whole stream of data) can be solved optimally, online detection is limited by the use of the subset of previous observations and reconsidering the whole history at each time step is computationally expensive. Second, detecting change points requires some prior belief on volatility, or hazard rate, *i.e.* the probability of a change point occurring at each time step. Living agents facing a new environment might not know this beforehand. Joshua Gold and colleagues provided a normative model for change-point detection parameterized by the hazard rate, H [179] that extends Bayes rule in the case of discriminating 2 sources. On each trial t, the prior belief for a source k writes :

$$\phi_{k,t} = \phi_{k,t-1} + \log\left(\frac{1-H}{H} + exp(-\phi_{k,t-1})\right) - \log\left(\frac{1-H}{H} + exp(\phi_{k,t-1})\right) \tag{1.21}$$

Where $\phi_k = \log(Pr\{\text{data comes from source } k\})$. They showed that human behavior conforms to the model predictions and that subjects adapt their subjective estimate of H to the actual hazard rate of the environment [179]. Several algorithms have been proposed to account for this adaptation, either via a hierarchical Bayesian model [180], an approximate Bayesian model that randomly samples values of H from a prior distribution [181], or a mixture of delta-rules with various learning rates [182].

An alternative approach is to suppose the regression towards a fixed prior ϕ_0 depending on a *forgetting factor* F [183] :

$$\phi_{k,t} = \log((1-F)\exp(\phi_{k,t-1}) + F\exp(\phi_0)) \tag{1.22}$$

This has the advantage of linearizing in the log-space and trivially expands to more than 2 sources. However, it relies on the assumption of equiprobability of the transitions between all sources, but offers the fixed prior ϕ_0 as a free parameter for arbitrary biases. The forgetting factor can itself be inferred using meta-forgetting factors in a hierarchical manner [183].

Experimental results consistently show that, in accordance with Bayesian models' predictions, subjects use larger learning rates after latent change points than during stable periods [184]. Moreover, they show that, in accordance with Bayesian inference, learning rates increase with the precision of the underlying generative distribution : the less noisy the observations, the larger the learning rates [184, 182]. Nevertheless, despite their starting point in the change-point detection formalism, these results are more compatible with continuous accounts of meta-learning described above in 1, than with Gallistel's radical position of learning as a feedforward process. Indeed, hierarchical Bayesian models describe feedback processes as they continuously need to update their beliefs at each level using bottom-up information. In addition, change points are identified probabilistically rather than deterministically, which conflates abrupt detection with fast, but continuous, learning rate adaptation. In fact, even a flat (non hierarchical) change-point detection model does not make behavioral predictions distinct from a hierarchical Bayesian model in simple reversal learning tasks [178].

Overlapping neural representations of continuous and discrete transitions

The use of simple learning tasks, such as probabilistic reversal learning or bandit tasks, might hinder the discriminability of continuous and discrete transitions at the neural level. Indeed, as we discussed at length in section 1, the tracking of unsigned prediction errors, trial-by-trial surprise or, more generally, volatility estimates are a key feature for smooth adaptation of learning parameters. A large body of evidence points towards the dorso-medial prefrontal cortex (dmPFC) and the adjacent dorsal anterior cingulate cortex (dACC) as a key hub for such computations [64, 74, 78, 81, 82, 185].

However, other studies suggest that cortical networks might undergo qualitative shifts during the course of learning, especially around latent change points [186, 187]. It seems that, when facing unexpected adverse events, subjects can abruptly switch to random strategies, most likely as an exploratory maneuver [77, 188, 104, 189, 190]. The locus coeruleus to dACC noradrenergic input has been thought to convey reset signals that trigger such random behavior [188, 104, 77, 190]. But beyond switching to random exploration states, activity in the dACC has been regularly associated with the selection of alternative strategies in general [80, 191, 95, 192, 193, 194, 195, 196].

Hence, opposite computational accounts have been mapped onto overlapping cortical regions, across rodents, non-human primates and humans. There are three possible, non mutually exclusive explanations for this discrepancy. First, prefrontal cortical neurons are known to display *mixed selectivity* [197]. Depending on the context, the same neurons could be tuned to adjust smoothly or to switch abruptly after error signals. Second, several authors proposed a fine-grained sub-regional specialization across the dmPFC [198, 199, 200]. Finally, the apparent paradox regarding the functions attributed to the dmPFC might emerge from inappropriate computational models and/or experimental tasks. Indeed, most of the models reviewed so far, whether they are meta-reinforcement learning models, hierarchical Bayesian filters or change-point detectors, are centered around the notion of optimality, *i.e.* how to adapt as quickly as possible to a changing world. This focus on the learning curve motivates parsimonious experimental paradigms, such as conditioning, reversal learning tasks or bandit tasks, in low-dimensional environments. But the purpose of implementing switching mechanisms to select appropriate strategies rather than optimizing the learning parameters of a unique strategy, goes beyond the acceleration of the learning curve. Comparing and selecting strategies allows agents to enrich their learning abilities through their knowledge of the underlying structure of the world. In the following section, I discuss models that employ such global meta-learning mechanisms.

2.2 Multimodular learning and cognitive flexibility

For over two decades now, many authors have suggested that (meta)learning processes arise from the interplay of multiple strategies or mixture of experts, that are selected depending on the inferred structure of the environment [201, 202, 203, 204, 205]. Thus, meta-learning is not only an optimisation mechanism but an organised set of functions handling abstract concepts. We will now examine three of them : model-based

learning (section 2.2), categorizaton (section 2.2) and episodic memory (section 2.2).

Model-based learning and mixture of experts

The associative models discussed in section 1 infer the latent structure of the environment only implicitly. In the field of reinforcement learning, one classically distinguishes **model-free** from **model-based** reinforcement learning [2]. Temporal-difference models belong to the former, since they do not explicitly infer or manipulate the underlying connections between states, but rather assume that temporal contiguity reveals structural proximity. Model-based reinforcement learning directly uses the transition matrix between states in order to infer the value of a state-action pair. In other words, on each trial the agent does not have access to a cached value for different options, but needs to plan a course of action given the transition matrix, or model :

$$Q_{t+1}^{MB}(s,a) = R_t(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{k \in \mathbf{A}} Q_t^{MB}(s',k) \tag{1.23}$$

With $R_t(s,a)$ and $T(s,a,s)$ the expected immediate return and transition probability from state s to state s' given action a respectively.

Since they have complementary benefits and costs, several authors have argued for the existence of both a model-free and a model-based learning systems in mammals [206, 207, 208, 209] (and a similar and earlier proposition by Nakahara, Doya and colleagues [210, 211]. While the model-free system has a constant cost but cannot plan ahead when changes occur, the model-based system has access to the underlying map of the environment and can quickly adjust to changes of the value function, at the expense of high deliberation costs. Switching from one to the other would therefore make the most of each system. Indeed, experimental data in rodents [209], non-human primates [212] and humans [208] showed that subjects tend to use a mixture of both strategies, or experts, in navigation tasks.

However, the nature of the arbitration process that would select model-based or model-free learning remains an open question. A natural suggestion, from the Bayesian point of view, is the relative uncertainty of both strategies [206, 213, 214]) : one of the experts would take over when it predicts future outcomes with a better accuracy than the other. In a reinforcement learning framework, without an explicit probabilistic representation of the expert's accuracy, one can use the squared prediction errors to compute the weights of each expert [215]. Other suggestions have been made, such as the speed/accuracy trade-off [216], or the average return [217]. Little is known about the neural implementation of the arbitration mechanism, though the lateral OFC has been proposed as a necessary structure for alternating between model-based and model-free learning in rodents [218].

The idea of mixing experts goes beyond the model-based/model-free dichotomy. For example, in addition to the two-expert mixture, a third expert has been proposed in order to account for animal behavior : a random exploratory strategy [219, 218]. This is equivalent to a dynamic adaptation of random exploration, depending on the reliability of the learning experts. Mixture of experts is a general approach for tackling learning in

complex environments, suggested early on from the structural segregation of sub-units in the striatum, striosomes, that could implement specialized policies [220, 202]. This has been formalized independently but with quite similar approaches by Gianluca Baldassare [202] and Kenji Doya's multiple model-based reinforcement learning (MMRL) [201]. MMRL is based on a mixture of model-based experts that partition the environment into spatial and temporal sub-domains. Each expert has its own value and transition function, hence a specific policy, and is weighted by its *responsibility signal*. The responsibility signal of the expert k at trial t, $\psi_t(k)$ corresponds to the posterior probability that the last observation is attributable to this expert. Note that this is a straightforward credit-assignment problem. Assuming a non-informative prior over the set of experts (*i.e.* they are all equiprobable) and normally distributed observations $\mathbf{x}$, the responsibility signal is written :

$$\psi_t(k) = \frac{e^{-\frac{1}{2\sigma^2}(x_t - \hat{x}_t(k))^2}}{\sum_i e^{-\frac{1}{2\sigma^2}(x_t - \hat{x}_t(i))^2}} \tag{1.24}$$

This is equivalent to the aforementioned softmaximization of squared prediction errors, with an inverse temperature of $-\frac{1}{2\sigma^2}$. Besides the potential specialization of each expert, mixtures also have the advantage to by-pass precise timing computations required by classical TD-RL without loosing the ability to anticipate rewards in complex environments [33, 203]. Indeed, reward anticipatory signals in the ventral striatum of rats have been shown to lack temporal precision [39] : In a task where rewards are delivered in sequences whose lengths varied according to a complex rule (location in a plus maze and visual cues), phasic excitatory activity in the VS was recorded before each delivery, but also after the last delivery, in violation of classical TD-RL. This result can be accounted for by a mixture of experts, each dealing with partial knowledge about the current state [33, 203, 39].

Hence, in addition to a faster detection of change points in a volatile environment, mixture of experts can optimize cost/benefit trade-offs and efficiently partition high dimensional state-spaces. However, the construction of experts remain a puzzling question, as straightforward weighting of experts from their own performance can hurt convergence towards efficient strategies [33, 203].

Categories and task-sets

Notwithstanding the above-mentioned work on the prefrontal cortex functions in (meta)learning, primary interest in those regions came from abstract reasoning, inductive logic and categorisation in perceptual decision making. Lesion studies in humans revealed the causal implication of the dorsolateral PFC for switching between abstract rules [221, 56] and single unit recordings in monkeys showed that neurons in this region respond to task-related variables independently from irrelevant dimensions [222, 223, 224]. Moreover, complex and instructed rules in perceptual categorization are associated with patterns of activity that combine simpler rules, still in the dlPFC [225]. Perceptual decision-making is conceptually different from learning, as rules are instructed and not

discovered, and feedbacks are usually deterministic. However, the handling of global sets of rules tied to a specific context, or **task-sets**, rather than separate low-level associative links, might be a fundamental component of learning in primates [222, 226, 227, 228].

Categorisation-like phenomena during learning manifest as one-shot learning [229, 230, 231], or generalization from a few samples to analogous but never encountered situations [232]. Formally, building unbounded categories from sequential observations corresponds to a Dirichlet process [233, 234]. A Dirichlet process is parameterized by a base distribution Π_0 and a concentration parameter α. For each observation, the prior probability for being assigned to an existing category c is $\frac{n_c}{\alpha+N-1}$, with n_c the number of observations in the category c and N the total number of observations. The observation can also be assigned to a new category drawn from Π_0 with probability $\frac{\alpha}{\alpha+N-1}$. The identity of the category then depends on its constituents, and defines the likelihood function of successive observations. Importantly, any new observation can affect the whole history of previous attributions, leading to an exponential cost of computation for online applications. Once again, unbounded Bayesian inference is itself intractable [149]. Hence, approximate algorithms have been proposed [235] that can mitigate some of the computational burden, but not all of it [236]. Dirichlet process models capture behavioral features of humans in categorization tasks, such as one-shot learning [231, 232], inferring causal structure [237] or generalizing previous knowledge to never encountered categories [238].

Learning by inferring the relevant task-set makes predictions regarding underlying neural activity. For example, abrupt behavioral transitions linked with context-related activity in the medial PFC have been found in rodents [239, 80] and non-human primates [240, 241]. In addition, BOLD signal in the mPFC is predictive of behavioral switches from spontaneous rule discovery in humans [242]. Taken together, these results suggest that the fundamental computational role of the mPFC is the appropriate selection of task-sets, thus encompassing previous interpretations in various contexts, such as volatility, exploration and conflict [61].

It has been suggested that task-sets are organized hierarchically following a rostro-caudal gradient in the prefrontal cortex [243, 244]. This gradient might reflect increasing abstraction levels [243, 245] or the temporal structure of the environment [244]. For example, in the *cascade model* [246], the lateral PFC is described as a hierarchy of controllers : action selection is the end result of nested control signals from the most proximal (sensory level) to the most temporally distal (branching). In a perceptual decision-making task, subjects were asked to apply an instructed rule based on cues with increasing dimensional complexity [246]. While BOLD activity in the premotor cortex was associated with sensory dimensions (dimensions directly associated with task relevant action), posterior lPFC activations correlated with contextual cues (irrelevant for the choice), and activity in the anterior lPFC reflected episodic cues (block-wide association rules) [246]. Moreover, an episode and its corresponding task-set, could itself be selected or interrupted at the highest level of the hierarchy, named *branching* control, supposedly implemented at the most anterior part of the lPFC, the frontopolar cortex [244].

Memory and episodic control

This hierarchical organization of the PFC underlies episodic control not only for instructed rules, but also for learned strategies. Indeed, a substantial part of learning effects in a multi-dimensional task are attributable to working memory rather than reinforcement learning *per se* [247]. Moreover, a hallmark of temporally structured environments is the recurrence of task-sets over time [248]. Koechlin and colleagues [249, 95] proposed a model for approximating Dirichlet processes using reinforcement learning to create multiple task-sets from a *probe* that acts as a base distribution using a long-term memory of previous episodes. In the *probe model*, possible task-sets are stored in a buffer, and the actor is selected according to its *reliability* for predicting future outcomes. Hence, recurrent rules do not need to be learned again. Each set is formalized as a Q-learning look-up table $\mathbf{Q}$ and a Bayesian look-up table $\mathbf{L}$ that stores the observed reward frequencies for each state-action pair. This allows to update the **reliability** $\Psi_{k,t}$ of each task-set $k \in [1, M]$, according to Bayes rule [249] :

$$\Psi_{k,t} \propto \Phi_{k,t} L_{k,t} \tag{1.25}$$

$$\Phi_{k,t+1} = \tau \Psi_{k,t} + \frac{1-\tau}{M-1}(1 - \Psi_{k,t}) \tag{1.26}$$

With $\tau \in [0, 1]$ the stability, *i.e.* the probability of non switching from trial t to trial $t+1$. Note that equation 1.26 gives an alternative parameterization of the prior probability compared to equations 1.21 and 1.22, assuming a uniform transition probability between sets. Moreover, in opposition to previously discussed mixture models, the selection of the acting set is binary : The task-set with a reliability above the threshold value of 0.5 takes control, and if no set reaches this criterion, the agent initializes a probe using its long-term memory of all the previously acting sets.

The probe model constitutes an approximate and tractable reduction of a Dirichlet process since updates of the latent variables only depend on the latest state (no backward pass over the whole history is required) and the space of possible sets is bounded by M. In addition, this model connects learning, categorization and episodic control since it can identify previously learned task-sets and directly reuse them. Human performance strongly supports this model over naive Q-learning in terms of learning speed and episodic memory [249, 95]. Moreover, neuroimaging and intra-cranial recordings reveal a mapping between prefrontal regions and specialized computations predicted by the model [95, 192] : While the medial PFC regions are associated with monitoring the reliability of the current actor and inferring episodic switches [192], BOLD signal in the lateral PFC correlates with the monitoring of alternative task-sets and set selection [95].

Hierarchical meta-learning models, such as MMRL, the cascade model or the probe model, crucially combine associative, local and gradual learning, with inferences over global and discrete states. This echoes recent suggestions in the field of artificial intelligence, that frame meta-learning as inferring the latent structure of the environment using past experience [131, 250, 251, 252]. This contrasts with approaches that define learning as

a domain-general associative process whose adaptations would require specialized sub-modules to control each degree of freedom (*e.g.* for volatility monitoring, attentional resources, scaling values, etc...). Interestingly, neural correlates of meta-learning processes tend to overlap in the same regions for quite different mechanisms. In the medial prefrontal cortex, ventral regions have been regularly associated with subjective values of elementary options [253] as well as the confidence in the subject's choices [254, 255] and the reliability of the actor task-set[95] or more generally abstract states forming a cognitive map [59]. Similarly, the dorsomedial PFC is typically involved in volatility monitoring [82], signalling surprise [64] or conflict [139] but also in abrupt switches towards exploratory strategies [193, 95, 256]. Finally, the lateral PFC often appears when studying counterfactual decision making [113], alternative exploratory strategies [89] or alternative task-sets [95]. Although all of these processes can co-exist as predefined hardwired mechanisms, it might be more parsimonious to suggest that specialised computations are selected according to their relevance to the task at hand. The central claim of this thesis is that meta-learning is a top-down process, that starts by identifying the space of relevant strategies before inferring the most appropriate. Furthermore, we hypothesize that global behavioral adaptations are ubiquitous and that several phenomena at a local level can be accounted for by top-down regulation processes.

3 Towards an unifying theory of meta-learning

3.1 Three axes of a meta-learning theory

In the previous review of existing meta-learning theories, I highlighted several inconsistencies both at the computational and neurophysiological levels. I now summarize these hypotheses along three axes, whose extremes are not necessarily mutually exclusive (they can coexist within separate modules), but whose integration remains unresolved : the direction (bottom-up vs. top-down), the scope (local vs. global) and and the continuity of adaptive processes.

Direction

Associative learning is fundamentally a bottom-up process. It was historically formulated in opposition to "*learning by having ideas*", *i.e.* involving high-level cognitive representations, and was operationalized by conditioning procedures using elementary stimuli and behavioral responses. Extending from this building block, meta-learning is thought of as a collection of peripheral modules to regulate the core process of forming associative links. Depending on the theoretical framework to which they belong, they may consist of an optimisation of reinforcement learning hyperparameters (*e.g.* [78]), a recursive hierarchy of Bayesian inference processes from the most local (*i.e.* perceptual) level (*e.g.* [151]) or a nested selection of task-sets from the most elementary level (*e.g.* [249]). Theories that give such a central place to a domain-general, automatic learning mechanism have been referred to as a dual-processing accounts, in opposition to purely top-down views [257].

While defending the prevalence of top-down processes in meta-learning (at least in humans), we do not want to claim that associative learning is a mere artifact of high-level cognitive processes. Associative learning remains a powerful theory, and recasting classical conditioning as a propositional reasoning might seem far fetched (see [257] and commentaries). However, there is compelling evidence to suggest that instrumental behavior is influenced by high-level cognitive control, for example via timescale adaptation, dimensionality reduction, counterfactual inference, combination of exploratory strategies or coordination of multiple task sets. Do these processes necessarily revolve around elementary low-level associative links ? If not, and instrumental behavior is the product of a flexible and abstract representation, then meta-learning is best described from the top down.

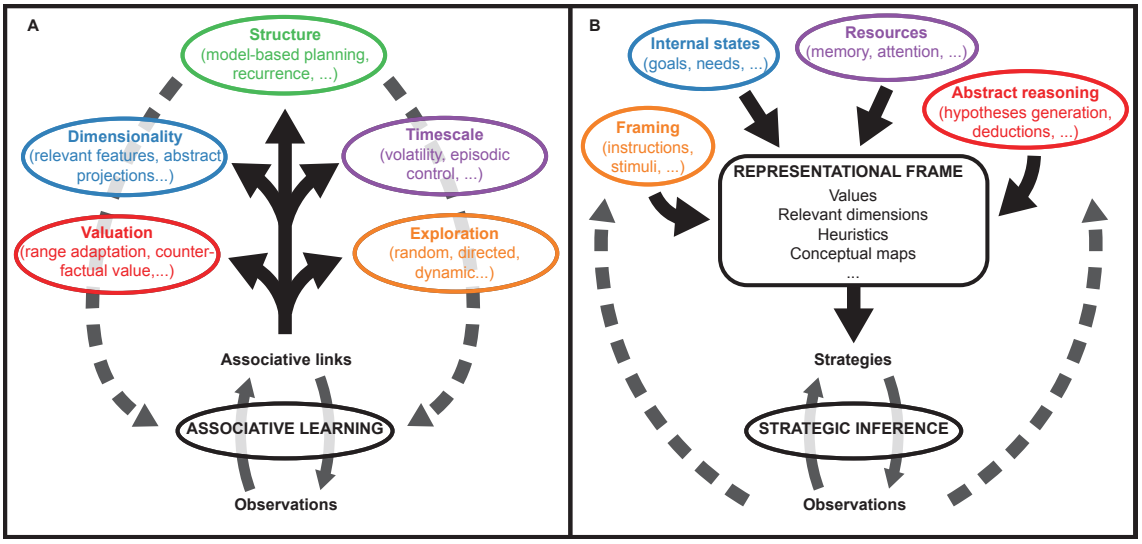

FIGURE *1.1* – ***Two perspectives on meta-learning A :*** *Meta-learning as a bottom-up process. Associative learning occupies a central position, and feeds higher-level specialized modules, via the by-products of local computations — e.g. anticipatory reward values, or prediction errors (thick, black arrows). These high-level processes can modulate the formation of associative links, as a retrocontrol (dashed, gray arrows).* ***B :*** *Meta-learning as a top-down process. Task-specific information is gathered from parallel high-level functions, to form a representational frame and generates a tailored strategic repertoire (thick, black arrows). Strategic inference operates at the level of these specialized strategies. High-level processes can feed from low-level computations or observations in order to adjust the representational frame (dashed, gray arrows).*

Scope

Another crucial and often overlooked aspect of meta-learning is the scope of the adaptive processes. For example, the environment's volatility might have different effects on learning whether it acts locally, *i.e.* on a single stimulus-action association, or globally, *i.e.* on a set of multiple associations [258]. In naturalistic environments, the underlying local state might not be as clear as in laboratory experiments, and learning is likely to

be affected by state uncertainty [39, 166]. Hence, the scope of (meta)learning might not only depend on the perceptual discriminability of elementary features *per se*, but on inferences about the most relevant level for the task at hand. For example, subjects can switch their representational basis from perceptual features to abstract objects depending on the predictive values of each format [259], or the naturalistic quality of stimuli [260].

Interestingly, this might explain the overlap of regions found to encode both local and global representations in the prefrontal cortex. For example, the vmPFC/OFC has been assigned many computational functions along the local-global axis : valuation of elementary choices [253], second-order judgements on choices (*i.e.* confidence)[255], global state posterior probability [261], or task-set reliability [95]. To reconcile these findings, it has recently been suggested that this region does not simply represent value (as other regions represent perceptual information), but builds it [262]. Therefore, the scope of meta-learning processes may not be absolute or predetermined, but dependent on the context, the qualitative characteristics of the external stimuli and the internal goals of the agent.

Continuity

The last line of debates regarding adaptive mechanisms is the opposition between a continuous view, where beliefs are gradually updated, and a discontinuous view, where agents search for change points between discrete entities [175, 186, 177]. For example, in accordance with the former view, it has been suggested that the learning rate could be adjusted depending on the volatility of rewards contingencies, in relation to activity in the dmPFC [82, 64, 78]. However, this region has been involved in abrupt switches in reversal learning [95] or foraging paradigms [193].

It is worth noting that this axis is still orthogonal to the previous one : global inference can result in continuous adaptations (*e.g.* if strategies are weighted by a responsibility signal [201]), or discrete switches corresponding to the selection of the most reliable task-set [249]. Moreover, discrete switches may not be distinguishable from fast continuous adjustments in simple tasks [178] and direct comparisons are still lacking [177]. Nonetheless, two points have to be clarified regarding this opposition. First, discontinuities in behavior do not imply discontinuities in the generative process. Agents might display abrupt behavioral transitions as a result of continuous but non linear adaptations. Discontinuities, whether through the selection of discrete objects or the crossing of thresholds, need to be tested thoroughly. Second, clear predictions from different models are required for this investigation to be successful. Indeed, in experimental paradigms where continuous or discrete modelling produce similar qualitative patterns, quantitative metrics alone provide less insight.

3.2 Hypotheses and predictions

Following these three axes, we derive the three general hypotheses regarding meta-learning :

1. Direction : Meta-learning relies on top-down dependencies, from high-level, abstract representations to elementary action selection. At the highest level, a **representational frame** is constructed in response to a specific problem. This encompass the goals, representational basis, relevant resources and perceptual dimensions to attend to. Within this representational frame, adaptive behavior emerge through **strategic inference**, *i.e.* selection of the most appropriate strategy.
2. Scope : Adaptations at a global level are ubiquitous and not peripheral to meta-learning. They emerge from strategic inference, in response to the latent structure of the environment, even when they are detrimental for the agent's performance.
3. Continuity : Whether surprise or volatility produce abrupt changes or continuous adjustments cannot be determined from the data, and is more a reflection of the model used for the analysis.

To test these hypotheses, we designed a probabilistic reversal learning task, involving several state-action associations, and varied the scope of latent contingency changes in different environments (for details see the Material and Methods section of the Featured Article). We make the following predictions :

1. Learning in complex (*i.e.* probabilistic, volatile and multidimensional) environments involves global mechanisms - **even when inappropriate or suboptimal**. Indeed, we predict model-free signatures of global adaptations by varying the scope of the true hidden changes.
2. Computational modeling of (meta)learning is critically dependant on the underlying representations strategic space. Comparing models that differ in the way they represent the problem at hand will favor the most specific formats (*i.e.* task-specific strategies) over generic ones (*i.e.* elementary associative links).
3. Some hyperparameters adaptations do not reflect active control, but are emergent features under the correct representational frame. In particular, modulation of internal variables of a model (*e.g.* the learning rate or the exploratory bias) might appear from strategic inference without requiring explicit implementation.
4. **Every adaptation can be fast** independently of how abstract they are. Due to the primacy of the representational frame, adaptations to high-order contextual statistics are observable at the timescale of our laboratory task.

Featured article

Recasting adaptability as strategic inference

Sami Beaumont[1], Mehdi Khamassi[2], and Philippe Domenech[1]

[1] *Neurophysiology of Repetitive Behavior, Institut du Cerveau, Sorbonne Université, Paris, France*
[2] *Institute of Intelligent Systems and Robotics, Centre National de la Recherche Scientifique, Sorbonne Université, Paris, France*

Abstract

Adaptive behavior is classically thought of as the expression of associative learning, *i.e.* the formation of stimulus-action links. However, higher-level control of learning processes, *i.e.* meta-learning, is not clearly understood. Here, we introduce a novel theory of adaptability in humans viewed as top-down high-level strategic inference rather than bottom-up low-level associative learning. We test this hypothesis in a modified Wisconsin Card Sorting Task, with multiple types of rule changes. Across five different contexts, we found consistent evidence that human subjects adapt to unexpected changes by manipulating global strategies, rather than learning local stimulus-action associations. Moreover, computational analyses of behavioral performance supported a model based on a predefined set of strategies, compared to bottom-up models that build strategies through trial and error. These results suggest that humans tend to reason at a strategic level and that local associative links might not be the necessary foundation for adaptability.

Introduction

In the early 1900s, Edward Thorndike's famous puzzle box experiment showed that complex strategies, involving multiple state-action associations, could be learned by trial and error in many species [1]. The idea of progressively learning an associative strength between cognitively meaningful events remains central in multiple fields, such as Psychology, Neuroscience and Artificial Intelligence. *Reinforcement learning* formalizes this process as a recursive update of state-action values by the error between observed reinforcements and its predicted values [2, 3]. This learning signal, called prediction error, has been related to phasic dopaminergic bursts broadcasted to cortical and subcortical networks, which offer a realistic neural implementation for this algorithm [4, 5, 6]. At the same time, however, studies on abstract categorization via card sorting tasks led to consider inference on global strategies (defined as sets of associations between stimuli and actions) as a crucial component of adaptability [7, 8, 9, 10].

The aim of the present study is to investigate the interplay between bottom-up associative learning and top-down strategic inference in humans. While the former mostly requires continuous adaptations based on local error signals, the later is characterized by discrete switches between global strategies.

Interestingly, both literature associates adaptability with overlapping sets of underlying brain regions. On the one hand, bandit and reversal learning tasks demonstrate fine-tuning of learning parameters to the uncertainty of expected outcomes [11, 12, 13], which is consistently associated with uncertainty monitoring in the dorso-medial prefrontal cortex (dmPFC) [11, 13, 14, 15]. Accordingly, mPFC core function has often been modeled as a meta-learner, dynamically adjusting associative learning to the latent statistics of the environment, using various algorithmic implementations such as a hierarchical Bayesian filter [16], an action-outcome predictor [15] or a reinforcement meta-learner [17, 18]. Alternatively, others argue that learning in volatile environments can be framed as detecting change points of the optimal policy [19, 20, 21]. Consistent with this theoretical framework, activity in the dmPFC reflects abrupt behavioral switches in rodents [22, 23, 24], non-human primates [25, 26, 27] and humans [28, 29], supporting the view that the mPFC implements a controller over abstract state-action rules, *i.e.* task-sets [30, 9, 31, 10].

Hence, while computational theories on human adaptability are divided into two opposite functions — the continuous adjustment of local learning versus the abrupt switches between high-level strategies — they consistently involve similar cortical regions, and in particular the dmPFC. This apparent paradox persists due to the lack of direct comparison of continuous versus change-point models in the literature [32] (although see [33]).

In this study, we present an original experimental paradigm allowing to measure global strategic inference in humans. By manipulating the types of covert rule changes in a modified Wisconsin Card Sorting task, we define behavioral markers of strategic inference. Using computational modelling, we investigate the extent to which associative learning and/or strategic inference explain human adaptability to abrupt changes. We hypothesize that adaptability is mostly the reflect of top-down strategic inference, inducing global effects that are ubiquitous in learning, even when unnecessary or harmful. This dependence on high-level abstractions such as behavioral strategies implies that efficient (meta)learning models must account for the underlying representation of the problem at stake (its *representational frame*). Consistent with this view, our results suggest that adaptability is primarily a reflection of strategic inference in our task, and that learning local associative links is not necessary. Furthermore, we show that continuous and discrete behavioral adaptations can emerge independently of the actual generative process and that the versatility of strategic inference may be responsible for the apparent paradoxes in the literature regarding the role of prefrontal sub-regions for meta-learning.

Results

Experimental paradigm : the Monkey feeding task

Healthy volunteers were required to perform a modified Wisconsin Card Sorting task, called the Monkey feeding task (figure 1). Participants had to find the correct fruit (banana, coconut or grapes) for each of 3 monkeys (different coat colors). At a given trial, they were shown one of the monkeys and could select one of the 3 fruits. Then, they received a feedback indicating if the monkey was happy or not with their choice. Importantly, participants were instructed that each monkey had only one preferred fruit at a time, that it could change overtime, and that one fruit could be correct for at most 2 monkeys. This limited the use of heuristics and deductive reasoning to infer correct associations. Finally, participants were instructed to maximize positive feedbacks, and were incentivized with a monetary bonus proportional to their performance. Feedbacks were reliable (*i.e.* matching the actual underlying rule) in 90% of the trials, thus participants had to learn the correct rules by trial and error over several trials. Moreover, the underlying fruit-monkey associations changed regularly during the task without notice. We call the latent structure of those changes an *environment*. A first group of participants underwent 2 environments over 4 sessions, and a second group 3 environments over 3 sessions (see Material and methods for details). Obviously, participants had no instructions regarding the different environments or conditions of the task. Finally, the order of environments in each experiment was counterbalanced across participants.

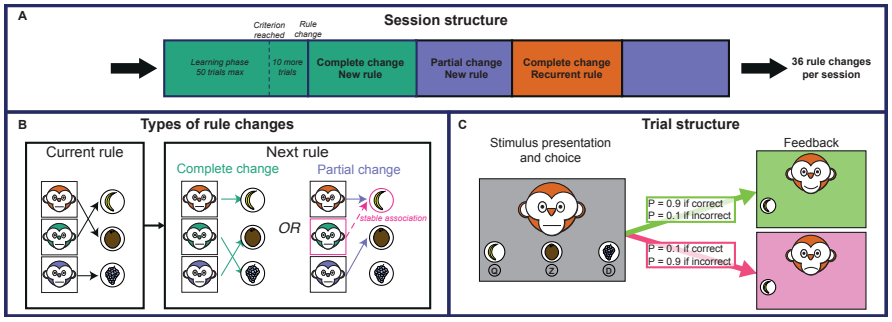

Figure 1: **The Monkey feeding task. A :** *During each session, participants had to learn associative rules by trial and error. Rule changes were triggered after subjects reached a performance criterion, followed by an asymptotic performance for 10 more trials.* **B :** *Rules changes could be either complete or partial (i.e. one association stayed the same). In the former case, the next rule could be the recurrent one.* **C :** *On each trial, subjects were presented one of the three monkeys, and had to choose one of the three fruit. Once they made a choice, a feedback screen indicated the monkey, the chosen fruit and the positive or negative reaction of the monkey. Crucially, only 90% of feedbacks reflected the actual rule.*

Behavioral signature of strategic inference

In the first environment, 72% of rule changes were complete, *i.e.* all monkeys changed their preferred fruit at the same time to form a new rule (16 changes) or went back to a (previously encountered) recurrent rule (10 changes). Some rule changes were partial (10 changes), meaning that only two stimulus-action-outcome associations changed and one stayed the same (which we refer to as the *stable association*). We reasoned that if subjects used inference at a global, strategic level, they would gain in performance for recurrent rules at the expense of their performance for partial rule changes. Indeed, partial rule changes would be detected later than complete rule changes, and could lead to global strategic switches, hurting the performance for the stable association. On the opposite, non-hierarchical associative learning would treat each association independently, thus the changing associations would be learned similarly in both complete and partial rule changes, while stable associations would be preserved.

In line with the strategic inference hypothesis, subjects were faster for learning recurrent rules compared to new rules (figure 2A-B). Moreover, to test the specificity of this effect to the repetition of a global rule and not the recurrence of local associative links, we compared the effect of new rules when constituted with rare associations (*i.e.* that had not been used for the previous 5 episodes at least) and new rules formed of more frequent associations. Both correct and exploratory responses were indistinguishable, suggesting that the frequency of local stimulus-action associations has no effect on memory (supplementary figure S2).

In addition, partial rule changes led to slower learning speeds (figure 2A) and a transient decrease of correct responses for the stable association (figure 2C), which we will subsequently refer to as the *interference effect*. Both effects were quantified by fitting a saturating exponential function with the learning curves for each subject in each condition (see Material and Methods). The recurrence effect was defined as the log-ratio between the slopes for recurrent and new rules (t-test against 0; $t_{50} = 2.12, p = 0.039$). The interference effect was the log-ratio between the asymptotic performance and the initial performance for stable associations (t-test against 0; $t_{50} = 8.89, p < 0.001$).

Identification of abrupt behavioral switches

Based on previous theoretical proposals, we investigated whether these recurrence and interference effects were the consequence of abrupt behavioral changes. To do so, we fitted a hidden Markov model of subjects' underlying strategy based on their choice patterns to find the position and the identity of each behavioral switch, for each subject (see Material and Methods). The model identifies the strategy most likely to be used on each trial, from among 27 possible strategies and a 28^{th} strategy corresponding to random behavior for every stimulus. Next, we classified strategic switches as *global* when the new strategy did not overlap with the previous one, *overlapping* when 1 or 2 associations were common to both strategies, or *random* when the new strategy is the random one.

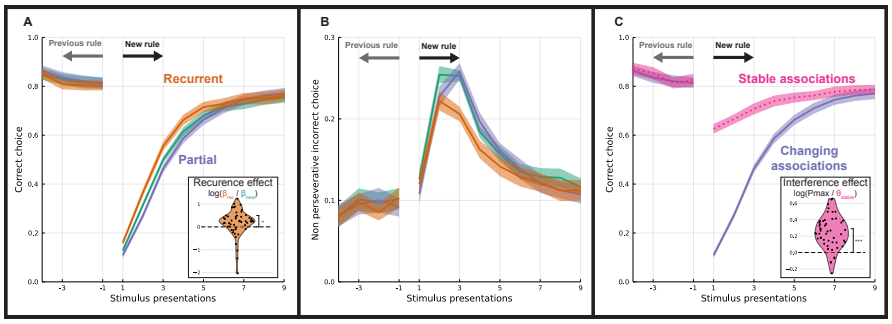

Figure 2: **Recurrence and interference effects** *: Behavioral performance in environment A1 (mean +/- sem).* **A :** *Participants learned faster after a rule change when the rule was recurrent, and slower when the change was partial. Insert : distribution of the recurrence effect (the log-ratio of the slope of the learning curve for the recurrent rule β_{rec} over the slope for new rules after complete changes β_{new})* **B :** *Participants made less non perseverative errors after a recurrent rule change.* **C :** *Performance decreased for stable associations after a rule change, even though the local contingency did not change. Insert : distribution of the global interference effect (the log-ratio between asymptotic performance $P_{\max}$ and initial performance fo stable associations θ_{stable})*

Perseverative responses locked on strategic switches revealed abrupt behavioral transitions (figure 3A). Remarkably, the decrease of performance for stable associations in partial rule changes was entirely explained by unwarranted non-overlapping switches (figure 3B). Moreover, strategic switches were likely inferred from underlying rule changes. First, global switches were more frequent after complete rule changes, whereas overlapping switches were over-represented after partial rule changes (figure 3C). Second, consistent with the idea that partial rule changes hinder inference at the strategic level, strategic switches of all types were more delayed after partial rule changes (figure 3D). We computed the latency from rule changes to strategic switches as the average number of stimulus presentations in-between. Using the number of presentations, and not trials, allowed to cancel out the effect of having a variable number of changing associations (2 or 3). Latencies were significantly higher after partial changes than complete changes for all switch types (Wilcoxon signed rank test; global : $Z = 603$, $p = 0.009$; overlapping : $Z = 924$, $p < 0.001$; random : $Z = 928$, $p < 0.001$). No significant differences were found when comparing latencies after the recurrent and new rule changes (global : $Z = 399$, $p = 0.052$; overlapping : $Z = 366$, $p < 0.833$; random : $Z = 422.5$, $p = 0.094$). Finally, the reward rates from rule changes to switches to the random strategy were below chance level, contrary to other strategic switches (figure 3D). This suggests that random behavior was not distributed equally during the task, but rather reflects information seeking in response to high uncertainty about the true contingencies.

As with other strategic switches, transitions in and out of random exploration were abrupt (figure 4A and B). Moreover, and consistently with the hy-

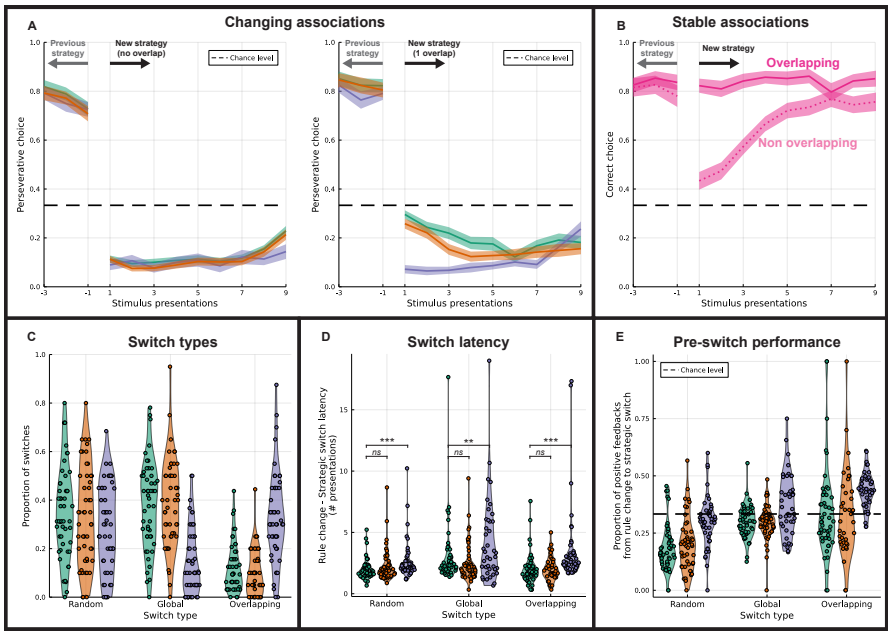

Figure 3: **Abrupt switches underlie associative learning** *: Hidden Markov model's output in environment A1.* **A :** *Perseverative responses (mean +/- sem) locked on global (left) and overlapping (right) strategic switches, after complete (green), recurrent (orange) and partial (purple) rule changes.* **B :** *Correct responses (mean +/- sem) for stable associations after partial rule changes locked on overlapping (solid line) and non overlapping (dotted line) strategic switches.* **C :** *Proportion of random, global and overlapping switches after a rule change.* **D :** *Switch latencies (average number of stimulus presentation) in environment A1. All strategic switches came later after a partial rule change than a complete rule change.* **E :** *Random switches were associated with below chance reward rates from the rule change, whereas other strategic switches were not.*

pothesis of an adaptive exploratory strategy, subjects used this strategy for fewer trials when the true rule was the recurrent one (figure 4C; $t_{46} = 2.27, p = 0.028$). This indicates that these bouts of random behavior were likely an active exploration strategy aimed at gathering evidence when the correct strategy could not be identified, rather than unspecific selection noise.

Computational modelling and model comparison

Our subjects displayed behavioral patterns compatible with strategic inference and abrupt behavioral transitions. However, similar patterns can be explained by associative models with a hierarchical control of hyperparameters. In order to discriminate between strategic inference and adaptive associative learning, we assessed five candidate models (see Material and Methods). Three of them were meta-adaptive Q-learners with dynamic learning rate, inverse temperature

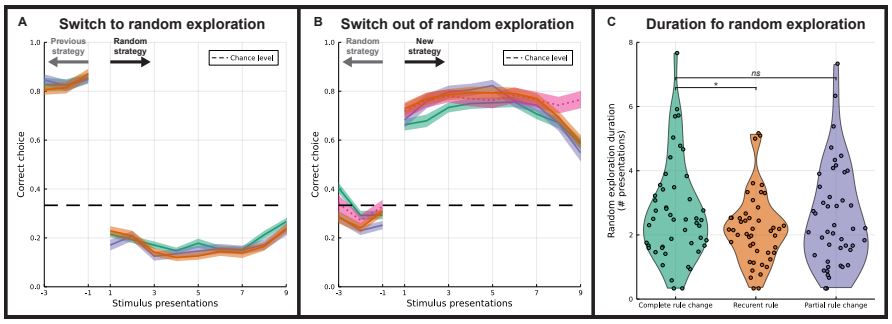

Figure 4: **Random behavior as an active exploration strategy** *:* **A-B :** *Perseverative (left) and correct (right) responses (mean +/- sem) locked on switches to and out of random exploration respectively.* **C :** *Random strategy duration (average number of stimuli presentations) in environment A1. Participants used the random strategy shorter when the rule changed to a recurrent rule.*

or both. We also tested the *Probe model* [28], that explicitly handles multiple task-sets built bottom-up using Q-learning. Finally, we designed an original model based on the assumption that strategic inference occurs within a pre-defined space of global stimulus-action mappings. We refer to this model as *Top-down strategic inference*, as it does not rely on bottom-up information to build a strategic repertoire.

Thus, our model has directly access to a complete repertoire of strategies, which comprises the $3^3 = 27$ possible stimulus-action mappings in our task. Strategies are selected based their reliability ϕ, which is updated according to Bayes rule :

$$\phi_{t+1}^k \propto \psi_t^k \ell_t^k \tag{1}$$

With ψ_t^k the prior probability of strategy k at trial t and ℓ_t^k the likelihood of the event at trial t given the strategy k : $\ell_t^k = \rho$ for expected events under strategy k and $\ell_t^k = 1 - \rho$ otherwise, with $\rho \in [0, 1]$ the evidence weight. The prior probability is computed using the forgetting factor framework [21] :

$$\psi_t^k = \omega \phi_t^k + (1 - \omega)\pi_t^k \tag{2}$$

With $\omega \in [0, 1]$ the forgetting factor, and π_t^k the hyperprior for strategy k at trial t. This hyperprior represents the global average of the reliability vectors over the course of an experimental session, *i.e.* a memory trace of past strategies. Crucially, strategies could compete for action selection only if they pass a threshold θ. If no strategy reaches this threshold, actions are selected randomly. Thus the probability of choosing action a is given by :

$$P(a) = \frac{\sum_k w_{a,k}\phi^k}{\sum_k w_{a,k}} + \frac{(1 - \sum_k w_{a,k})}{3} \tag{3}$$

The left part of the sum being the marginal evidence for action a over the set of strategies that pass the threshold. Note that this set is defined probabilistically

with a weighting vector $\mathbf{w}$:

$$w_{a,k} = Pr\{\phi_k > \theta\}\delta(a,k) \tag{4}$$

With δ the Dirac function such that $\delta(a,k) = 1$ if action a is prescribed by the strategy k and 0 otherwise. Using $Pr\{\phi_k > \theta\}$ instead of an all or nothing thresholding allows to account for stochasticity in the strategic selection, and facilitate inference for model fitting (see Material and Methods).

This model captured all the behavioral features of our subjects in this first experiment (figure 5A-C). It also performed closer to humans than any other models in terms of learning efficiency (figure 5D) and global effects (figure 5E). Furthermore, predictive accuracy, computed by 6-fold cross validation (see Material and Methods), favored the top-down strategic inference model over bottom-up models (figure 6).

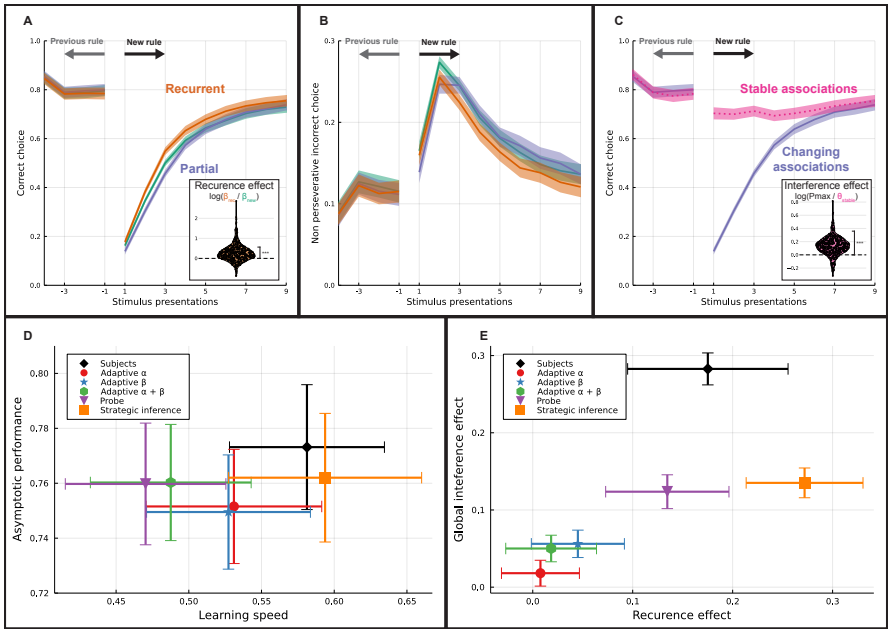

Figure 5: **Top-down strategic inference captures behavioral features better than bottom-up associative models** *:* **A-C :** *Simulations of the strategic inference model reproduce main qualitative patterns of human subjects in environment A1.* **D-E :** *Qualitative model comparison along 4 dimensions. Strategic inference is closer to human subjects in learning speed, asymptotic performance and the recurrence effect. The interference effect is less well captured though no other model gets closer.*

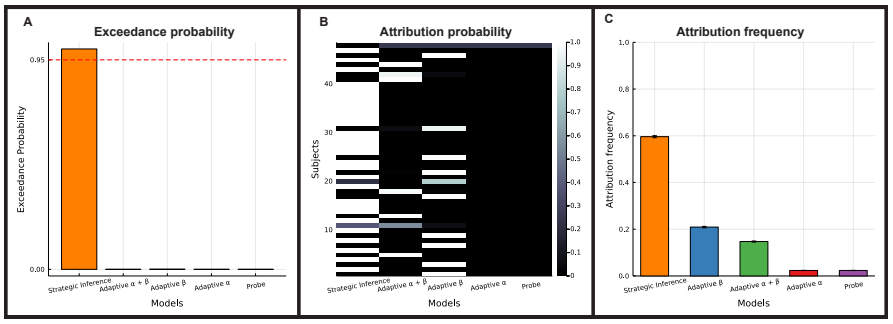

Figure 6: **Quantitative model comparison favors top-down strategic inference over bottom-up associative models** *: Quantitative model comparison using 6-fold cross validation.* **A :** *Exceedance probability of strategic inference compared to other models is close to 1.* **B-C :** *Predictive accuracy favors strategic inference in more than 60% of subjects.*

Continuous and discrete meta-adaptations as emergent features

Importantly, fitting the hidden Markov model on simulated data revealed that abrupt transitions were not sufficient to rule out associative learning models. Indeed, as shown in figure 7, even an adaptive Q-learning model could display clear cut behavioral switches. Nevertheless, the distribution of strategic switches lacked the use of active bouts of random exploration exhibited by human subjects (figure 7B, right panel). By contrast, the strategic inference model showed a similar pattern to that of our subjects (figure 7A, right panel).

On the other hand, the strategic inference model can behave as an associative learning model with adaptive hyperparameters. Figure 8 shows the effect of internal variables on the model's exploratory bias and apparent learning rate. Due to the thresholding imposed on strategy selection, action probabilities not only depend on the marginal evidence (*i.e.* the marginal reliability of all strategies that prescribe a specific action), but also on the entropy within the space of strategies (*i.e.* the spread of those reliabilities). Furthermore, the update of the marginal evidence does not monotonically grow with the prediction error (the difference between the observed outcome and the prior marginal evidence). This is due to the use of Bayes rule to update the reliabilities : very surprising events can induce inversely correlated updates because of low prior beliefs (figure 8B-C). Interestingly, high forgetting tends to linearize the updates (figure 8C) at the expanse of slower convergence toward the correct strategy, due to high volatility.

Contextual adaptation of strategic inference

The same participants also performed 2 sessions out of 4 in another variant of the task, without any cues or instructions indicating the differences. In this environment, all rule changes were partial, some of them with two changing

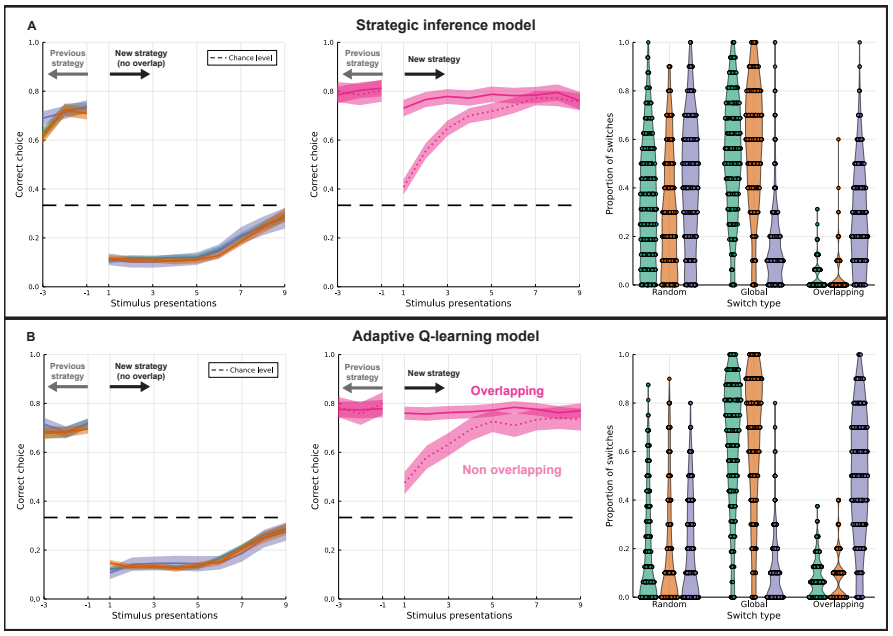

Figure 7: **The use of a random exploratory strategy, but not abrupt behavioral switches, is a marker of strategic inference. A :** *Fitting the hidden Markov model on simulated data from the strategic inference model revealed abrupt transitions (left and middle panel) and a distribution of strategic switches (right panel) comparable to human subjects.* **B :** *Using simulated data from an adaptive Q-learning model showed similar abrupt transitions, but missed the use of random exploratory strategies.*

associations (*2/3 rule change*) and some with one changing association (*1/3 rule change*). In half of the latter, feedback reliability transiently decreased to 75% around change points. Importantly, this increased feedback noise was only applied to stable associations, so the local information required to learn the new rule was comparable across conditions. Finally, a similar noise increase could transiently be applied to 2 associations out of 3 at a distance from the rule changes, as a control.

Although there was no complete change in this environment, it was still possible to identify global inference effects (figure 9). Participants learned new associations faster and made less exploratory choices when changes occurred simultaneously for 2 associations compared to isolated changes (figure 9A). Moreover, the performance for stable associations tended to be lower after 2/3 rule changes than isolated changes (figure 9B). Similarly, adding noise to stable associations' feedbacks led to more exploratory choices for changing stimuli (figure 9C) and performance for stable associations was more strongly impacted by noise and rule changes than noise alone (figure 9D). Taken together, those results suggests that subjects still relied on strategic inference despite an environ-

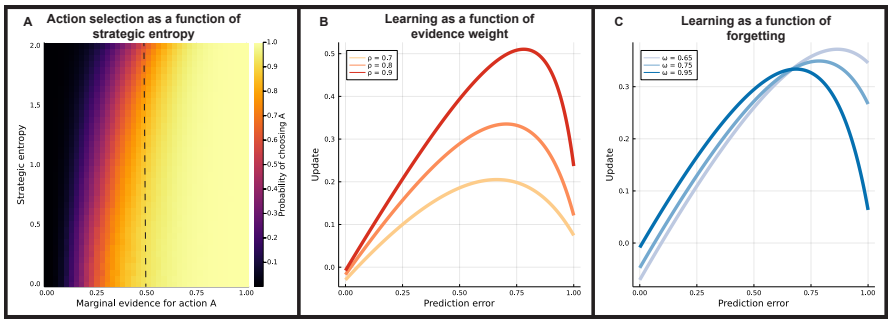

Figure 8: **Action selection and learning emergent adaptations in the strategic inference model** *:* **A :** *Probability of choosing an action as a function of the marginal evidence in its favor and the strategic entropy. Action selection tend to maximize evidence when the entropy is low.* **B-C :** *Updating of marginal evidence towards an action as a function of the difference between the observed reward and the current marginal evidence, with various values for parameters ρ and ω. The evidence weight ρ controls the size of the update, while the forgetting factor ω controls the linearity of the update for high prediction errors.*

ment promoting local associative links over global rule changes. Consistently, all these behavioral patterns are reproduced by the strategic inference model (supplementary figure S1).

Furthermore, comparing both environments revealed a striking difference in the subjects' behavior. Figures 10A and B show that learning speed and interference effect were much more pronounced in the first environment, despite similar rule changes. This indicates that participants adapted to session-wide contextual statistics : the presence of complete rule changes prompted more radical (figure 10C) and faster (figure 10D) strategic switches. More specifically, after a 2/3 rule change (a condition common to both environments), the proportion of random and global switches were significantly higher in the first environment ($t_{49} = 2.27$, $p = 0.027$; $t_{49} = 2.17$, $p = 0.035$ respectively; overlapping : $t_{49} = -0.85$, $p = 0.397$), and the latency was significantly shorter for switches to the random strategy ($Z = 707$, $p < 0.001$) but not the others (global : $Z = 230$, $p = 0.331$; overlapping : $Z = 565$, $p = 0.082$). This effect is not attributable to proximal differences in the rate of positive feedbacks since, once again, the conditions were exactly similar in both environments and the pre-switch performance (figure 10E) was comparable for random and global switches ($t_{50} = 0.22$, $p = 0.825$ and $t_{49} = -0.031$, $p = 0.976$ respectively). However, subjects engaged into overlapping switches after observing slightly lower reward rates in the second environment ($t_{50} = -2.36, p = 0.022$).

Splitting the experimental sessions in thirds (supplementary figure S3), revealed that these differences emerged quite rapidly, and were already perceptible during the first epoch (13 initial episodes). Such fast adaptations might be a direct consequence of the frequency of complete rule changes in each environment. However an alternative explanation is that subjects adapted to a lower local volatility in the second environment. Indeed, in this environment the ab-

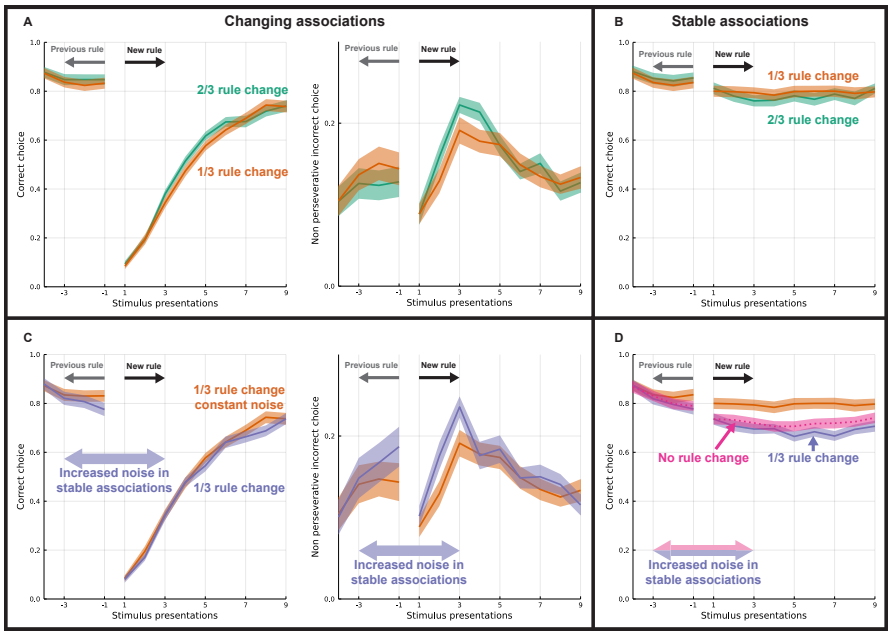

Figure 9: **Global effects remain in a less structured environment** *: Behavioral performance in environment B1 (mean +/- sem).* **A :** *Participants tended to learn faster and made less non perseverative errors when the rule changed for 2/3 of the associations than when it changed only for 1/3.* **C :** *Participants tended to make more errors when learning new associations while the stable associations receive noisier feedbacks.* **B & D :** *The global interference effect tended to be more pronounced when more associations changed.*

sence of complete changes made local associations more stable, which could have led our participants to be more conservative. Interestingly, when comparing the values of fitted free parameters of the strategic inference model in both environment, only two of them were statistically different (figure 11) : the evidence weight ($t_{50} = 3.11, p = 0.003$) and the average threshold for strategic selection ($t_{50} = 3.51, p = 0.001$). This computational distinction might reflect complementary mechanisms. On one hand, reduced local volatility could lead to lower evidence weight, as each individual observation carried less information on a potential rule change. One the other hand, a lower threshold allows to consider more strategies for action selection, which is convenient when successive rules tend to overlap.

Control study

To further investigate the effects of strategic inference and its contextual modulation, we designed a complementary study for a second group of participants. The aims of this control study were : 1/ to independently replicate both recurrent and interference effects, and 2/ to test the local volatility hypothesis

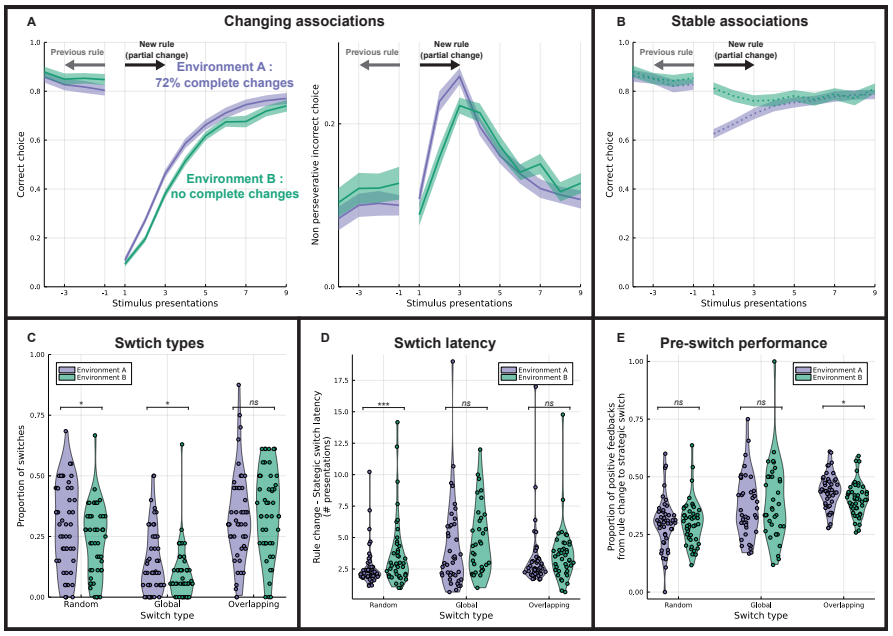

Figure 10: **Global effects are modulated by covert contextual statistics** *: Comparison between similar rule changes in environments A1 and B1.* **A-B :** *Behavioral performance after 2/3 partial rule changes in both environment. Participants learned faster and displayed a more pronounced global interference in the environment where rule changes were mostly complete (environment A1, purple) compared to the environment when all rule changes were partial (environment B1, green).* **C :** *When rule changes were mostly complete (environment A1), participants displayed a bias towards random and global strategic switches.* **D :** *When all rule changes were partial (environment B1) random strategic switches were more delayed.* **E :** *Pre-switch performance did not differ between both environments.*

for contextual modulation. In our initial paradigm, the use of memory was assessed by the recurrence of the same rule over the course of a session. This might enhance performance either via episodic memory of global rule sets, or via local repetition effects of the same associations. In the control study, we used 3 recurrent rules that were intertwined with new rules during one session. Another session contained either complete or partial (2/3) rule changes, to replicate the interference effect. Finally, participants had to perform a third session, with only partial rule changes, on 1 or 2 associations. Again, no instructions or cues were given regarding the differences between sessions, and their order was counterbalanced across subjects. Crucially, in all three sessions, the local rate of changes was matched : whether changes were complete or partial, each local rule for a stimulus was changed every 14.5 trials on average. This allowed to control for local volatility as a possible confounding factor for contextual modulation.

Figure 12 shows that both the recurrence ($t_{54} = 2.85$, $p = 0.006$) and the

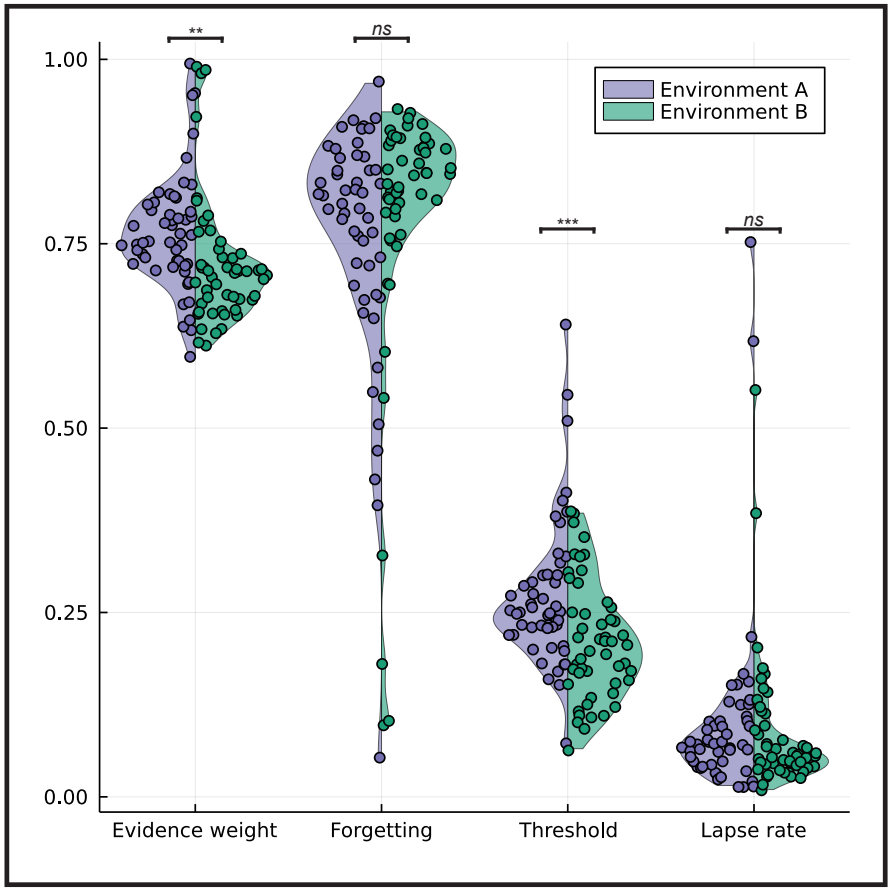

Figure 11: **Computational dissection of the contextual modulation** *: Within-subject comparison of fitted free parameters in environment A1 and B1. The top-down strategic inference model accounts for slower learning and attenuated global interference effect in environment B1 by decreasing the trial-by-trial evidence weight and strategic threshold.*

interference effects ($t_{54} = 6.93$, $p < 0.001$) were reproduced. Importantly, the recurrence effect was not significantly different in the first study, with only one recurrent rule, compared to the control study with 3 recurrent rules (Two-sample t-test; $t_{104} = -0.66$, $p = 0.509$). This confirms that subjects tended to reason over global strategies rather than local associative links in our task. However the results regarding contextual adaptation were more mixed : while learning speed remained slightly faster in environment A than in environment B, the interference effect on stable associations was of similar magnitude (figures 12C and 13B, $t_{54} = 1.18$, $p = 0.24$). This might be due to a reduced proportion of complete changes in the control study (45%) compared to the original study (72%).

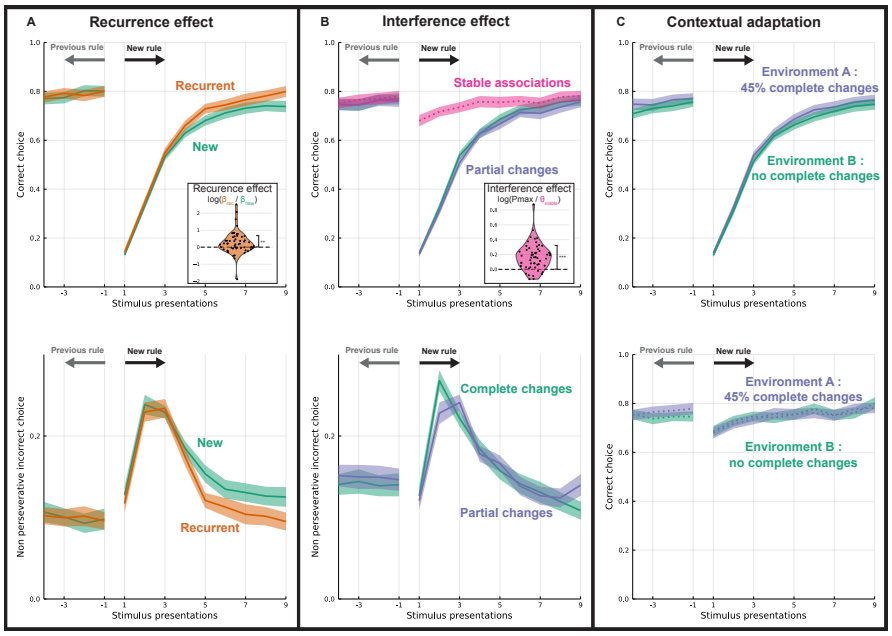

Figure 12: **Reproducibility of global and contextual effects** *: Behavioral results of the control study.* **A :** *Reproduction of the recurence effect with 3 recurrent rules instead of only one.* **B :** *Reproduction of the global interference effect.* **C :** *Contextual modulation of global effects reduced after matching for local frequency of changes.*

In accordance with this idea, figure 13B shows that reducing the proportion of complete changes was associated with a statistically significant diminution of the interference effect (Two-sample t-test; $t_{103} = 2.02$, $p = 0.046$), while matching local volatility was indeed associated with a significant increase of the effect ($t_{104} = 4.82$, $p < 0.001$).

Discussion

In this work, we hypothesized that, in environments requiring to track sets of multiple stimulus-action associations, adaptability is better described as top-down strategic inference rather than bottom-up meta-learning, even with an explicit hierarchical representation of the task structure. Our behavioral and computational modelling results support this hypothesis. In particular, two model-free effects constituted the behavioral signature of strategic inference. First, we showed that humans were sensitive to the recurrence of previously encountered rules. This result is consistent with previous reports [28, 34], but we provide clear evidence that this recurrence effect is better accounted for by a memory of global strategies rather than local associative links. Indeed, the effect size did not significantly vary with the number of recurrent rules,

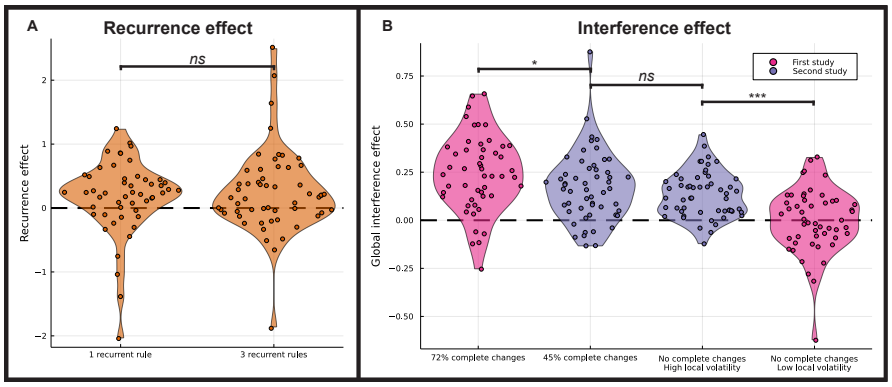

Figure 13: **Global effects across environments** *: Recurrence and interference effects in the 2 studies.* **A :** *Th recurrence effect was of similar magnitude in both studies, despite the different number of recurrent rules.* **B :** *The intereference effect varied as a function of the proportion of complete changes and local volatility.*

and, in the first study, learning efficiency was not affected by the frequency of local associations (supplementary figure S2). This translates well in our model, where memory is represented by an hyperprior over reliabilities at the strategic level. Hence, only global rules will leave a distinguishable memory print, and increasing the number of recurrent sets does not necessary require more resources. Second, our subjects displayed non-trivial interference patterns after partial rule changes. While this effect could be accounted for by bottom-up models (*e.g.* using a dynamic inverse temperature or abrupt switches in Q-learning models), only the top-down strategic inference model displayed both qualitative and quantitative behavioral similarities to our participants.

Meta-learning as a top-down process

The idea that learning relies on detecting hidden changes and selecting appropriate strategies resonates with Gallistel's famous hypothesis of the learning curve as an artifact of abrupt detections averaged over a population [19]. It is also consistent with recent findings regarding structured strategic shifts across species [24, 23, 26, 29, 35, 36]. Nevertheless, detection of behavioral switches with optimal detectors such as a hidden Markov model should not be taken as definitive evidence against gradual associative learning. In fact, we show that such methods would detect abrupt transitions in behavior, as soon as learning is fast enough with typical non linear action selection functions. Thus it is the type of those transitions, in particular global random exploration strategies, their frequency and delay that offer distinctive features of strategic inference against associative learning.

Moreover, our results demonstrate the necessity for high-level adaptive mechanisms, as subjects displayed dramatic variations across environments, despite

similar local conditions. This is at odds with suggestions that that low-level noisy updates could account for near-optimal adaptability, without any explicit volatility estimates [37]. Indeed, in our task, subjects displayed quite different behavioral patterns after partial rule changes depending on the general context they were in, though factors that could modulate low-level noise, such as pre-switch performance, feedback reliability, and even the number of changing associations, were identical in both environments.

Explicit computational modelling of such an adaptability is currently lacking in our model. However, fitted values of free parameters and the results of the control study oriented towards two possible explanations for contextual modulation. First, subjects displayed variable evidence weighting, which seemed to correlate with local volatility. This manifested itself in slower learning in the more stable contexts. This is consistent with previous findings showing that the learning rate tend to follow the volatility of the environment [11, 18], and normative Bayesian accounts of adaptive learning [16, 38]. Crucially, our approach helped to disambiguate local from global volatility, revealing that matching local volatility could reduce the gap in learning speeds. In the strategic inference model, this makes sense as more stable local associations are less informative regarding strategic changes, thus lowering the global evidence weight. Second, the threshold at which strategies are considered for action selection was also affected by environmental statistics : with more partial rule changes, subjects displayed lower threshold values. This could be explained by the need for considering more contiguous strategies (and thus maintaining lower thresholds) when external changes are more overlapping. Consistently we found that the global interference effect, a model-free marker of strategic shifts, varied with the proportion of partial changes in the environment.

Taken together, our results suggest that meta-learning is better described as a top-down process, starting from high-level representations of the problem at hand, drawing the contours of appropriate behavioral strategies. This does not mean that low-level associative learning never plays a role, nor that all learning is reducible to propositional reasoning as some authors have suggested [39]. Associative learning is a powerful framework, and the articulation between low-level mechanisms and strategic inference would require additional studies. Nevertheless, we agree with previous proposals that place learning at the intersection of other cognitive functions, such as detection [19], categorization [40, 41] or memory [42, 43]. In other words, high-level representations of the task at hand form a *representational frame* that conditions downstream learning. This representational frame may itself be inferred, by induction or deduction, but may also be sensitive to non-inferential processes such as external instructions or idiosyncratic hypotheses generation.

The importance of the representational frame

Placing meta-learning under the constraints of a representational frame accounts for the diversity of learning mechanisms in the literature. For example, previous work on navigation showed that humans can learn from a mixture of model

free and model based strategies [44, 45]. Switching from one mode to the other is itself a non trivial adaptation problem, and several authors have proposed arbitration mechanisms depending on computational constraints and expected gains of both strategies [46, 47, 48]. Interestingly, recent studies showed that explicit knowledge of task structure (through instructions to the subjects) might enhance model-based inference [49, 50]. Moreover, it has been suggested that usual measures of the model-free/model-based balance might be biased by memory of previous policies [51]. Those results consistently show that behavioral features in learning tasks are not the product of isolated learning mechanisms, but are strongly influenced by other high-level functions.

In fact, one could argue that building a representational frame prior to the task is a necessary condition for efficient learning in complex environments. Indeed, associative learning requires fine-tuning of contextual hyperparameters, such as range adaptation [52], counterfactual inference [53] or identification of relevant features [54]. It is worth noting that the first two follow naturally from inference on global policies, since selection of the best strategy does not require any absolute valuation of individual options. This does not imply that such valuation never occurs, as empirical results clearly demonstrate that subjects remain sensitive to value differences in post-learning tests [52]. However, there is no definitive evidence that range adaptation is a purely bottom-up process. Similarly, identification of relevant features and attentional orientation have been framed as bottom-up processes, built upon local prediction errors [54]. A recent report showed that, while selection of the subset of relevant features increased with value gathering during the course of learning, it rests on abstract representations of alternative objects [55]. Online shaping of the representational frame surely occurs, for example attention and memory can be oriented via local associative signals (*e.g.* prediction errors) [56]. However these low-level computations require dimensionality reduction beforehand to be tractable, thus they cannot be sufficient for representation learning.

Hence, though low-level information can adjust the representational frame, our results are in line with previous proposals that insist on the inherent limits of such bottom-up processes to account for human behavior [57, 58]. However, understanding how high-level representations are put in place remains a puzzling question. This relates to the notion of, compositionality *i.e.* the ability to generate new concepts by combining sub-parts of existing ones [59], for example through hierarchical Bayesian inference [60, 61], though alternative algorithms have been suggested [62, 63, 64]. Interestingly, compositional generalization is often presented as a specific ability of primates [65], and the question remains whether top-down strategic inference within a flexible representational frame can be applied to learning in other animal models, such as rodents. In this respect, further research could provide a better understanding of the respective prevalence of low-level associative mechanisms and top-down processes in learning, as well as the role of the underlying neural structures.

Material and Methods

Participants

The study was approved by the Sorbonne Université ethics committee. All participants were recruited via e-mail and gave their informed consent before participation in the study. Subjects were screened for the absence of any history of neurological and psychiatric disease or any current psychiatric medication.

Main study : 58 subjects connected at least once, and 51 subjects (31 female, age 33 ± 15) completed the 4 sessions and were included in the analysis. They were paid a fixed amount of 40€ plus a maximum bonus of 20€ proporitional to their overall performance (mean: 15 ± 4.7€).

Control study : 74 subjects connected at least once, and 55 (42 female, age 29 ± 13) completed the 3 sessions and were included in the analysis. They were paid a fixed amount of 30€ plus a maximum bonus of 15€ proporitional to their overall performance (mean: 8 ± 5.6€).

Experimental tasks

On the first day of the study, subjects were sent a link valid during two weeks to perform the tasks online. The tasks were programmed in JavaScript using the jsPsych framework, and all the scripts, data and links were supported by JATOS. The main study was hosted on MindProbe.eu, the server of the European Society for Cognitive Psychology (ESCoP), while the replication study was hosted on the local Brain Institute (ICM) server in Paris.

Main study : On each trial, subjects were asked to select one of the three fruits (grapes, banana or coconut) to feed one of the three monkeys (green, blue and orange) currently displayed on the screen. After fruit selection, a feedback screen appeared to indicate if the answer was correct (the monkey liked the fruit) or incorrect (the monkey did not like the fruit). Crucially, in order to maintain some degree of uncertainty on the latent rules underlying the task, monkeys' fruit preference was stochastic : in 90% of the trials, a monkey would like only its preferred fruit (that the subjects had to learn) and dislike the two other fruits, while the remaining 10% of the trials consisted in trap trials where monkeys disliked their favorite fruit and liked the two others.

Subjects were informed that each monkey had only one preferred fruit at a given time, but one fruit could be preferred by at most 2 monkeys. Subjects were instructed that each monkey would occasionally change their preferred fruit without any cue. There was no instruction on the structure of such changes (*i.e.,* subjects were not biased to believe that all 3 monkeys would change their preferred fruit at the same time).

Each subject had to complete 4 sessions of 39 episodes. Each episode lasted 20 to 60 trials. When subjects reached a performance criterion, the episode lasted 10 trials more before ending. The performance criterion was to get 4 correct answers out of the last 5 trials and 2 consecutive correct answers for each monkey. The rule change for the new episode was one of 4 possible conditions,

depending on the latent structure of the session :

Environment A1:

1. **Complete new rule** (12 episodes): All the monkeys changed their preferred fruit at the same time, forming a new mapping. Most of the time these mappings were never encountered before, but in order to solve the combinatorial problem, a few rules were considered new when not encountered for more than 8 episodes (*i.e.* more than approx. 160 trials).

2. **Partial new rule** (10 episodes): Two out of three monkeys changed their preferred fruit, forming a new mapping.

3. **Recurrent rule** (10 episodes): All the monkeys changed their preferred fruit, forming the same recurrent mapping than one already encountered before by the subject.

4. **Rare rule** (4 episodes): All the monkeys changed their preferred fruit, forming a new mapping. Not only this mapping was never encountered before, but it was formed by associations that were unused for more than at leat 5 episodes (*i.e.,* more that approx. 100 trials). Since no difference was found in the subjects' performance between this condition and the complete new rule condition, we merged them in the main analyses.

Environment B1:

1. **2/3 rule change** (9 episodes): Two out of three monkeys changed their preferred fruit. This condition is similar to the *Partial new rule* condition in Context A.

2. **1/3 rule change** (9 episodes): Only one monkey changed its preferred fruit.

3. **1/3 rule change + noise** (9 episodes): One monkey changed its preferred fruit, but for the two others, the probability of trap trials (*i.e.,* inacurate feedbacks) increased to 25% from ten trials before the partial rule change to ten trials after.

4. **No rule change + noise** (9 episodes): None of the monkeys changed their preferred fruit, but the probability of trap trials increased to 25% for two monkeys, from ten trials before the partial rule change to ten trials after.

The 4 conditions were balanced within blocks of 13 episodes. At the end of a block, subjects had a 2-minute break. The next block started with the previous block's last episode.

The order of the 2 environments was balanced between subjects: half of them did 2 sessions of Environment A1 first, the other half started with 2 sessions of Environment B1.

On each trial, subjects had 3 seconds to answer, after which the trial was considered invalid. After 5 consecutive invalid trials a warning appeared on the screen. After a second warning, the whole session ended and was considered invalid.

Second experiment (control study) : The instructions, the structure of trials and blocks, were similar to the main study. However, in order to match the global frequency of local rule changes, the length of the entire session was kept constant at 1564 trials, and the duration of each episode varied depending on the condition. The total number of local changes for each context was 108. The participants had to perform 3 sessions, each with a different context, with different conditions:

Environment A2 :

1. **Complete new rule** (18 episodes - 39 to 45 trials each): All the monkeys changed their preferred fruit, forming a new mapping.
2. **Recurrent rule** (18 episodes - 39 to 45 trials each): All the monkeys changed their preferred fruit, forming one of 3 recurrent mappings previously encountered.

Environment A3 :

1. **Complete new rule** (20 episodes - 34 to 38 trials each).
2. **Partial (2/3) new rule** (24 episodes - 31 to 35 trials each): Two out of three monkeys changed their preferred fruit.

Environment B2 :
(NB : In this environment, since local changes are more intertwined, the number of trials per episode was such that one monkey would not change its preferred fruit twice in less than 30 trials. Thus, even if the global rule set often changed, local associations were maintained long enough for the task to be manageable.)

1. **2/3 new rule** (36 episodes): Two out of three monkeys changed their preferred fruit.
2. **1/3 new rule** (36 episodes): Only one monkey changed its preferred fruit.

Fitting of the learning curves

We quantified the learning effects by fitting a function of the performance as the number of stimulus presentation from the rule changes :

$$f(n) = \theta + (p_{\max} - \theta) * (1 - e^{-\beta(n-1)}) \tag{5}$$

With $0 \leq p_{\max} \leq 1$ the asymptotic performance, $0 \leq \theta \leq 1$ the performance for $n = 1$ and $\beta \geq 0$ the learning slope. We let β vary with the condition, and θ had different values for stable and changing associations. This model was fitted for each subject, with an adaptive Hamiltonian Monte Carlo scheme [66].

Identification of global behavioral switches

In order to detect global behavioral switches from the dataset, we made the assumption that at any trial each subject would act according to a strategy, *i.e.,* a stimulus-response mapping or *task-set*. Thus, we used a hidden Markov model (HMM), with the underlying strategy of the subject as the hidden variable X. Since there is 3 stimuli and 3 possible choices in our task, the set of all the possible strategies has a cardinality of 27. We added a 28th strategy corresponding to random choices for all 3 stimuli. Hence the vector $\mathbf{X}$ of all strategies is indexed from 0 to 27, with X_0 the random strategy

Parametrization of the HMM: We assumed a constant emission probability, *i.e.,* the probability of acting according to the underlying strategy and not randomly, $0 \leq \rho \leq 1$. The transition probabilities between strategies were parameterized by a vector $\mathbf{\Theta}$ such that :

$$\theta_i = Pr\{Dist(X^+, X^-) = i\},\ 0 \leq i \leq 3 \tag{6}$$

$$\theta_4 = Pr\{X^+ = X_0\} \tag{7}$$

With X^+ and X^- the next and current strategy respectively, and $Dist(X^+, X^-)$ the distance between strategies. The distance was simply the number of differences between the 2 mappings, hence θ_0 corresponds to the probability of keeping the same strategy between 2 consecutive trials, while θ_3 is the probability of a global behavioral switch. Finally, the free parameter $0 \leq \tau \leq 1$ controlled the probability of staying in the random strategy.

HMM learning: HMM learning corresponds to the procedure of finding the most likely hyperparameters given the data. In our model there were 7 hyperparameters, ρ, $\mathbf{\Theta}$ and τ. We used MCMC samples from the posterior distribution of the hyperparameters for each experimental session and each subject. The log-likelihood of the hyperparameters given the sequence of stimuli $\mathbf{S}$ and choices $\mathbf{A}$ is:

$$\mathcal{L} = \sum_t \log\left(\sum_k \frac{\pi_{k,t}\ell_{k,t}}{Z_t}\right) \tag{8}$$

where $\pi_{k,t}$ is the log prior probability of the strategy k on trial t, $\ell_{k,t}$ the log-likelihood for strategy k on trial t and Z_t the normalizing constant. Thus:

$$\pi_{k,t} = \sum_{j=1}^{27} \theta_{Dist(j,k)}\pi_{j,t-1} + \frac{1-\tau}{27}\pi_{0,t-1}$$

$$\pi_{0,t} = \sum_{j=1}^{27} \theta_4\pi_{j,t-1} + \tau\pi_{0,t-1}$$

$$\ell_{k,t} = \begin{cases} \rho, & \text{if } A_t \text{ is the choice mapped with } S_t \text{ under strategy } k \\ 1-\rho, & \text{otherwise} \end{cases}$$

with $\pi_{k,0} = -log(28)$ for all k.

We used a fully adaptive Hamiltonian Monte Carlo scheme [66] to sample from the posterior distribution of the hyperparameters.

HMM inference and global switches detection: HMM inference refers to the process of infering the most likely sequence of hidden variables given a sequence of observations and a set of hyperparameters. With our samples of the hyperparameters posterior, we could generate samples of the posterior over the hidden variables, *i.e.,* the strategies used on a trial-by-trial basis, using Viterbi algorithm. Then, we were able to identify for every episode the trial with the highest probability of switch, and the identity of such a switch (switch to a random strategy, global strategic switch or overlapping strategic switch).

Generative models

Counterfactual Q-learning

Four models were variants of a basic counterfactual Q-learning model, in which Q-values where udpated as :

$$Q_t(a,s) = \begin{cases} Q_{t-1}(a,s) + \alpha(r_t - Q_{t-1}(a,s)), & \text{if a is the chosen action} \\ Q_{t-1}(a,s) + \kappa\alpha(1 - r_t - Q_{t-1}(a,s)), & \text{otherwise} \end{cases} \tag{9}$$

With $0 \leq \alpha \leq 1$ the factual learning rate and $0 \leq \kappa \leq 1$ the counterfactual learning rate. This parametrization insures that the counterfactual learning rate is always lower than the factual one, but follows the same dynamics (in the case of an adaptive learning rate).

Actions were sampled from a semi-uniform Boltzman (softmax) distribution :

$$Pr\{a|s\} = (1-\epsilon)\frac{e^{\beta Q(a,s)}}{\sum_b e^{\beta Q(b,s)}} + \frac{\epsilon}{3} \tag{10}$$

With $\beta \geq 0$ the inverse temperature, and $0 \leq \epsilon \leq 1$ the lapse rate.

Adaptive learning rate

The first variant featured an adaptive learning rate as in the classical Pearce-Hall model [13] :

$$\alpha_t = \alpha_{t-1} + \eta(|\delta_t| - \alpha_{t-1}) \tag{11}$$

With $|\delta_t|$ the absolute value of the (factual) prediction error at trial t and $0 \leq \eta \leq 1$ a meta-learning rate.

Adaptive inverse temperature

In the second variant the inverse temperature β was adapted to the dynamics of the average reward rate, as in [67] :

$$\beta_t = \beta_{t-1} + \lambda(\bar{\bar{R}}_t - \bar{R}_t) \tag{12}$$

$$\bar{R}_t = \bar{R}_{t-1} + \eta(r_t - \bar{R}_{t-1}) \tag{13}$$

$$\bar{\bar{R}}_t = \bar{R}_{t-1} + \eta(\bar{R}_{t-1} - \bar{\bar{R}}_{t-1}) \tag{14}$$

With $\bar{R}_t$ and $\bar{\bar{R}}_t$ the first and second order moving averages of the reward rate respecitvely, η a meta-learning rate and $\lambda \geq 0$ a scaling factor.

Probe model

The last variant using Q-learning was the Probe model [28, 34]. For details about this model we refer to [28]. We made four modifications to the Probe model to fit our experimental design. First, the Q-learning module was updated as described above, contrary to the original model where counterfactual learning did not have a free parameter. Second, we removed inter-stimuli counterfactual learning, as in our task a response can be correct for multiple stimuli. Third, we removed the entropy-based computation of the probe's initial reliability, in favor of a free parameter for probe initialization. We made this change for fitting convenience as we found that the original parametrization led to bad convergence. Finally, we fitted the model with only 1 task-set at a time (in addition to the probe) for identifiability purpose, since in our first experiment there was only one recurrent rule and the long-term memory would already be biased by it. Thus, there was no need for monitoring more task-sets.

Top down strategic inference model

The strategic inference model can be seen as the generative counterpart of the hidden Markov model. Indeed, the 27 possible strategies were monitored by a corresponding reliability vector $\mathbf{\Phi}$ updated according to equations 1 and 2. Actions were sampled using a "soft" thresholding of the strategies : for each strategy, the probability of passing the threshold was used as a weight for computing action probability :

$$Pr\{a|s\} = (1-\epsilon)P(a,s) + \frac{\epsilon}{3} \tag{15}$$

$$P(a,s) = \frac{\sum_k w_{a,s,k}\phi^k}{\sum_k w_{a,s,k}} + \frac{(1-\sum_k w_{a,s,k})}{3} \tag{16}$$

$$w_{a,s,k} = Pr\{\phi_k > \theta\}\delta([a,s],k) \tag{17}$$

With δ the Dirac function such that $\delta([a,s],k) = 1$ if the strategy k comprises the association between action a and stimulus s, and 0 otherwise. $Pr\{\phi_k > \theta\}$ was computed using a probabilistic threshold $\theta \sim \text{Beta}(b_1, b_2)$. Hence :

$$Pr\{\phi_k > \theta\} = \text{B}(\phi_k; b_1, b_2) \tag{18}$$

With $\mathrm{B}(\phi_k; b_1, b_2)$ the incomplete beta function evaluated at ϕ_k. If all reliabilities for defined strategies were below the threshold, choices were selected randomly, which corresponds to the rightmost part of equation 16.

Model fitting and model comparison

Models were fitted using the MCMC package Turing for Julia 1.6 [68]. All the models, except the Probe model, were fitted using an adaptive version of Hamiltonian Monte Carlo [66]. The Probe model could not, due to discontinuities in the likelihood function because of task set selection and probe creation and deletion. Thus, we used a Gibbs sampling scheme, separating continuous parameters (learning rates, inverse temperature and lapse rate) that were fit using Hamiltonian Monte Carlo, from discontinuous parameters (volatility, probe's initial reliability and memory weight), that were fit using a slice sampling scheme. For all models, convergence was assessed by qualitative inspection of the chains, and checks of the $\hat{r}$ and ESS statistics.

For model comparison we derived the predictive accuracy of each model via a 6-fold cross validation scheme. For each subject, each of the two sessions was divided in 3 blocks of equal length that were used as testing datasets. The training datasets for each block comprised all the blocks of the other session, plus the previous blocks (if any) of the same session. Indeed, due to the temporal dependency of the latent variables, the models had to be fitted to up to the testing block. We then derived the expected log-predictive density as defined in [69], which was used to compute attribution frequencies and exceedance probabilities [70].

References

[1] Edward L Thorndike. *Animal Intelligence : experimental studies*. New York, 1911.

[2] Robert A Rescorla and Allan R Wagner. A theory of Pavlovian conditioning : The effectiveness of reinforcement and non- reinforcement. In *Classical Conditioning II : Current Theory and Research*, number January 1972. 1972.

[3] Christopher J C H Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3):279–292, 1992.

[4] W. Schultz, P. Dayan, and P. R. Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997.

[5] John O'Doherty, Peter Dayan, Johannes Schultz, Ralf Deichmann, Karl Friston, and Raymond J. Dolan. Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning. *Science*, 304(5669):452–454, 2004.

[6] Samuel J. Gershman and Naoshige Uchida. Believing in dopamine. *Nature Reviews Neuroscience*, 20(11):703–714, 2019.

[7] Esta A Berg. A simple objective technique for measuring flexibility in thinking. *The Journal of general psychology*, 39(1):15–22, 1948.

[8] Brenda Milner. Effects of different brain lesions on card sorting. *Archives of neurology*, 9(1):90, 1963.

[9] Katsuyuki Sakai. Task set and prefrontal cortex. *Annual Review of Neuroscience*, 31:219–245, 2008.

[10] Philippe Domenech and Etienne Koechlin. Executive control and decision-making in the prefrontal cortex. *Current Opinion in Behavioral Sciences*, 1:101–106, 2015.

[11] Timothy E.J. Behrens, Mark W. Woolrich, Mark E. Walton, and Matthew F.S. Rushworth. Learning the value of information in an uncertain world. *Nature Neuroscience*, 10(9):1214–1221, 2007.

[12] Elise Payzan-Lenestour and Peter Bossaerts. Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Computational Biology*, 7(1), 2011.

[13] Matthew R Roesch, Guillem R Esber, Jian Li, Nathaniel D Daw, and Geoffrey Schoenbaum. Surprise! neural correlates of pearce–hall and rescorla–wagner coexist within the brain. *European Journal of Neuroscience*, 35(7):1190–1200, 2012.

[14] Benjamin Y. Hayden, Sarah R. Heilbronner, John M. Pearson, and Michael L. Platt. Surprise signals in anterior cingulate cortex: Neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. *Journal of Neuroscience*, 31(11):4178–4187, 2011.

[15] William H. Alexander and Joshua W. Brown. Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, 14(10):1338–1344, 2011.

[16] Christoph Mathys, Jean Daunizeau, Karl J. Friston, and Klaas E. Stephan. A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, 5(MAY):9, 2011.

[17] Mehdi Khamassi, Pierre Enel, Peter Ford Dominey, and Emmanuel Procyk. Medial prefrontal cortex and the adaptive regulation of reinforcement learning parameters. In *Progress in Brain Research*. 2013.

[18] Massimo Silvetti, Eliana Vassena, Elger Abrahamse, and Tom Verguts. Dorsal anterior cingulate-midbrain ensemble as a reinforcement meta-learner. *PLoS computational biology,*, 14(8):e1006370, 2018.

[19] C. R. Gallistel, Terence A. Mark, Adam Philip King, and P. E. Latham. The rat approximates an ideal detector of changes in rates of reward: Implications for the law of effect. *Journal of Experimental Psychology: Animal Behavior Processes*, 27(4):354–372, 2001.

[20] Matthew R. Nassar, Joseph T. McGuire, Harrison Ritz, and Joseph W. Kable. Dissociable forms of uncertainty-driven representational change across the human brain. *Journal of Neuroscience*, 39(9):1688–1698, 2019.

[21] Vincent Moens and Alexandre Zénon. Learning and forgetting using reinforced Bayesian change detection. *PLoS Computational Biology*, 15(4):1–41, 2019.

[22] Dougal G.R. Tervo, Mikhail Proskurin, Maxim Manakov, Mayank Kabra, Alison Vollmer, Kristin Branson, and Alla Y Karpova. Behavioral variability through stochastic choice and its gating by anterior cingulate cortex. *Cell*, 159(1):21–32, 2014.

[23] D Gowanlock R Tervo, Elena Kuleshova, Maxim Manakov, Mikhail Proskurin, Mattias Karlsson, Andy Lustig, Reza Behnam, and Alla Y Karpova. The anterior cingulate cortex directs exploration of alternative strategies. *Neuron*, 109(11):1876–1887.e6, 2021.

[24] Daniel Durstewitz, Nicole M. Vittoz, Stan B. Floresco, and Jeremy K. Seamans. Abrupt transitions between prefrontal neural ensemble states accompany behavioral transitions during rule learning. *Neuron*, 66(3):438–448, 2010.

[25] René Quilodran, Marie Rothé, and Emmanuel Procyk. Behavioral Shifts and Action Valuation in the Anterior Cingulate Cortex. *Neuron*, 57(2):314–325, 2008.

[26] Yoshiya Matsuzaka, Tetsuya Akiyama, Jun Tanji, and Hajime Mushiake. Neuronal activity in the primate dorsomedial prefrontal cortex contributes to strategic selection of response tactics. *Proceedings of the National Academy of Sciences of the United States of America*, 109(12):4633–4638, 2012.

[27] A. Saez, M. Rigotti, S. Ostojic, S. Fusi, and C. D. Salzman. Abstract Context Representations in Primate Amygdala and Prefrontal Cortex. *Neuron*, 87(4):869–881, 2015.

[28] Anne Collins and Etienne Koechlin. Reasoning, learning, and creativity: Frontal lobe function and human decision-making. *PLoS Biology*, 10(3), 2012.

[29] Philippe Domenech, Sylvain Rheims, and Etienne Koechlin. Neural mechanisms resolving exploitation-exploration dilemmas in the medial prefrontal cortex. *Science*, 369(6507):eabb0184, aug 2020.

[30] Earl K Miller and Jonathan D Cohen. An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci*, 24:167–202, 2001.

[31] David Badre, Andrew S. Kayser, and Mark D'Esposito. Frontal cortex and the discovery of abstract action rules. *Neuron*, 66(2):315–326, 2010.

[32] Dimitrije Marković and Stefan J. Kiebel. Comparative analysis of behavioral models for adaptive learning in changing environments. *Frontiers in Computational Neuroscience*, 10(APR), 2016.

[33] Micha Heilbron and Florent Meyniel. Confidence resets reveal hierarchical adaptive learning in humans. *PLoS Computational Biology*, 15(4):1–24, 2019.

[34] Mael Donoso, Anne G.E. Collins, and Etienne Koechlin. Foundations of human reasoning in the prefrontal cortex. *Science*, 344(6191):1481–1486, 2014.

[35] Zoe C Ashwood, Nicholas A Roy, Iris R Stone, Anne E Urai, Anne K Churchland, Alexandre Pouget, and Jonathan W Pillow. Mice alternate between discrete strategies during perceptual decision-making. *Nature Neuroscience*, 25(2):201–212, 2022.

[36] Maïlys C.M. Faraut, Emmanuel Procyk, and Charles R.E. Wilson. Learning to learn about uncertain feedback. *Learning and Memory*, 23(2):90–98, 2016.

[37] Charles Findling, Nicolas Chopin, and Etienne Koechlin. Imprecise neural computations as a source of adaptive behaviour in volatile environments. *Nature Human Behaviour*, 5(1):99–112, 2021.

[38] Payam Piray and Nathaniel D. Daw. A simple model for learning in volatile environments. *PLoS Computational Biology*, 16(7):1–26, 2020.

[39] Chris J Mitchell, Jan De Houwer, and Peter F Lovibond. The propositional nature of human associative learning. *Behavioral and Brain Sciences*, 32(2):183–198, 2009.

[40] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.

[41] Sang Wan Lee, John P. O'Doherty, and Shinsuke Shimojo. Neural Computations Mediating One-Shot Learning in the Human Brain. *PLoS Biology*, 13(4):1–36, 2015.

[42] Anne G E Collins and Michael J Frank. How much of reinforcement learning is working memory , not reinforcement learning ? A behavioral , computational , and neurogenetic analysis. *European Journal of Neuroscience*, 35(December 2011):1024–1035, 2012.

[43] S. Ritter, J. X. Wang, Z. Kurth-Nelson, and M. Botvinick. Episodic Control as Meta-Reinforcement Learning. *bioRxiv*, pages 948–953, 2018.

[44] 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.

[45] 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(OCTOBER 2012):1–19, 2012.

[46] Wouter Kool, Fiery A. Cushman, and Samuel J. Gershman. When Does Model-Based Control Pay Off? *PLoS Computational Biology*, 12(8):1–34, 2016.

[47] Marios C. Panayi, Mehdi Khamassi, and Simon Killcross. The rodent lateral orbitofrontal cortex as an arbitrator selecting between model-based and model-free learning systems. *Behavioral Neuroscience*, 135(2):226–244, 2021.

[48] Rémi Dromnell, Erwan Renaudo, Guillaume Pourcel, Raja Chatila Benoît, and Girard Mehdi 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. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12413 LNAI:68–79, 2021.

[49] Carolina Feher da Silva and Todd A. Hare. Humans primarily use model-based inference in the two-stage task. *Nature Human Behaviour*, 4(10):1053–1066, 2020.

[50] Pedro Castro-Rodrigues, Thomas Akam, Ivar Snorasson, Marta Camacho, Vitor Paixão, Ana Maia, J Bernardo Barahona-Corrêa, Peter Dayan, H Blair Simpson, and Rui M Costa. Explicit knowledge of task structure is a primary determinant of human model-based action. *Nature Human Behaviour*, pages 1–16, 2022.

[51] Oliver Vikbladh, Daphna Shohamy, and Nathaniel D. Daw. Episodic Contributions to Model - Based Reinforcement Learning. *Cognitive Computational Neuroscience Conference*, pages 2016–2017, 2017.

[52] Stefano Palminteri and Maël Lebreton. Context-dependent outcome encoding in human reinforcement learning. *Current Opinion in Behavioral Sciences*, 41:144–151, 2021.

[53] Erie D. Boorman, Timothy E. Behrens, and Matthew F. Rushworth. Counterfactual choice and learning in a Neural Network centered on human lateral frontopolar cortex. *PLoS Biology*, 9(6), 2011.

[54] Yuan Chang Leong, Angela Radulescu, Reka Daniel, Vivian DeWoskin, and Yael Niv. Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments. *Neuron*, 93(2):451–463, 2017.

[55] Aurelio Cortese, Asuka Yamamoto, Maryam Hashemzadeh, Pradyumna Sepulveda, Mitsuo Kawato, and Benedetto De Martino. Value signals guide abstraction during learning. *eLife*, 10:1–27, 2021.

[56] Angela Radulescu, Yeon Soon Shin, and Yael Niv. Human Representation Learning. *Annual Review of Neuroscience*, 44:253–273, 2021.

[57] Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:1–72, 2017.

[58] Pedro A. Tsividis, Joao Loula, Jake Burga, Nathan Foss, Andres Campero, Thomas Pouncy, Samuel J. Gershman, and Joshua B. Tenenbaum. Human-Level Reinforcement Learning through Theory-Based Modeling, Exploration, and Planning. *arXiv preprint arXiv*, pages 1–67, 2021.

[59] Dedre Gentner. Structure-Mapping: A Theoretical Framework for Analogy. *Cognitive Science*, 7:155–170, 1983.

[60] Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[61] Jane X. Wang. Meta-learning in natural and artificial intelligence. *Current Opinion in Behavioral Sciences*, 38:90–95, 2021.

[62] Daniel J. Navarro and Amy F. Perfors. Hypothesis Generation, Sparse Categories, and the Positive Test Strategy. *Psychological Review*, 118(1):120–134, 2011.

[63] Mark Blokpoel, Todd Wareham, Pim Haselager, Ivan Toni, and Iris van Rooij. Deep Analogical Inference as the Origin of Hypotheses. *The Journal of Problem Solving*, 11(1), 2019.

[64] Ronald B. Dekker, Fabian Otto, and Christopher Summerfield. Curriculum learning for human compositional generalization. *Proceedings of the National Academy of Sciences of the United States of America*, 119(41):1–12, 2022.

[65] Aldo Genovesio, Steven P Wise, and Richard E Passingham. Prefrontal–parietal function: from foraging to foresight. *Trends in cognitive sciences*, 18(2):72–81, 2014.

[66] Matthew D Hoffman, Andrew Gelman, et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

[67] 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.

[68] Hong Ge, Kai Xu, and Zoubin Ghahramani. Turing: a language for flexible probabilistic inference. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, pages 1682–1690, 2018.

[69] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432, 2017.

[70] Klaas Enno Stephan, Will D Penny, Jean Daunizeau, Rosalyn J Moran, and Karl J Friston. Bayesian model selection for group studies. *Neuroimage*, 46(4):1004–1017, 2009.

Supplementary material

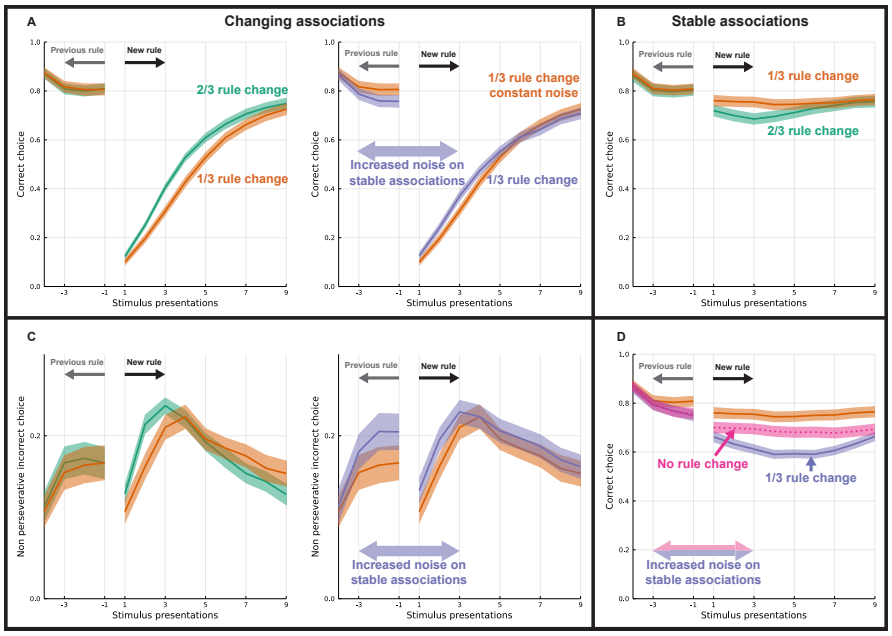

Figure S1: **Global effects in environment B1 are all predicted by the top-down strategic inference model** *: Simulations of the strategic inference model reproduce main qualitative patterns of human subjects in environment B1.*

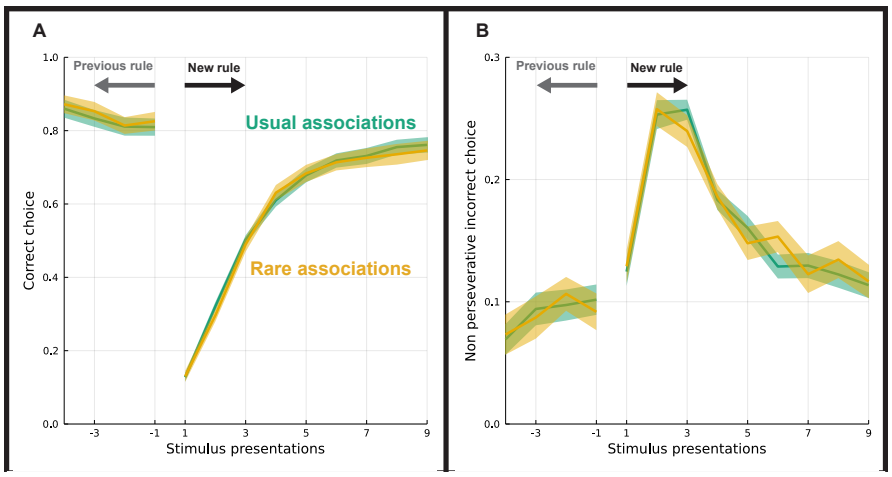

Figure S2: **No effect of local recurrence** *: Correct (A) and exploratory (B) responses for complete rule changes, formed with usual associations (green) or rare associations (yellow). Rare associations were defined as associations that have not been correct during the previous 5 episodes.*

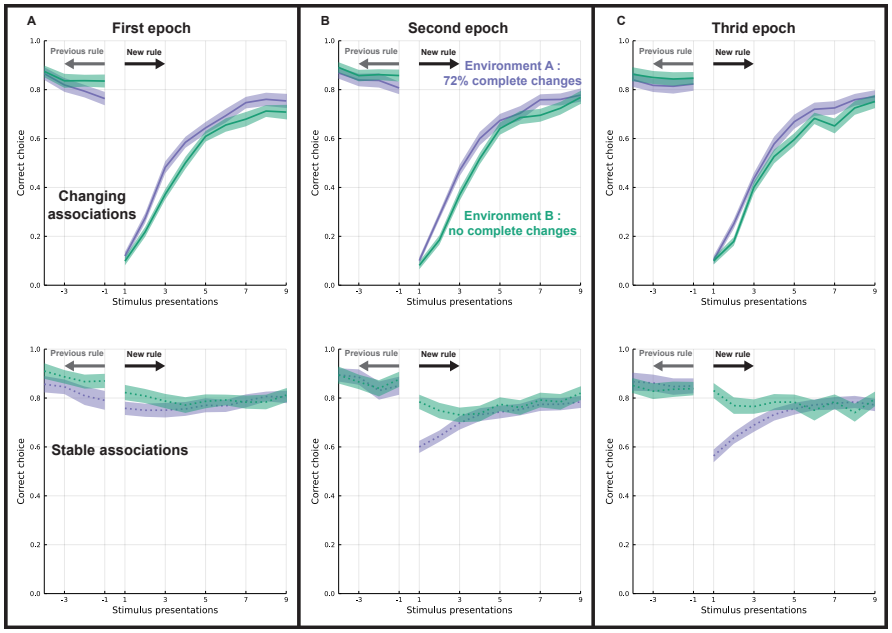

Figure S3: **Interference effect over the course of an experimental session** *: Correct responses for changing (top row) and stable associations (bottom row), splitted by session's epoch.*

General discussion

1 The representational frame of meta-learning

Our results suggest that, in tasks involving several stimulus-action associations, recasting (meta)learning as strategic inference provides a better account of human performance than bottom-up associative learning models. Moreover, our subjects demonstrated dramatic variations of their learning and exploration patterns in relation to covert statistical regularities in the environment. We believe that such versatility is a marker of top-down processes relying on prior delineation of the the space of concepts relevant to the task : for instance the goals, the temporal dynamics, the causal structure of the environment or the space of available strategies. All these elements constitute a **representational frame**. While inference based on bottom-up information processing is necessary to adjust this representational frame, other mechanisms may play a critical role to build it. The relationship between low-level associations formation and high-level cognitive processes, and their relative importance for human learning, has been debated for a long time (for a recent review see [257], and the commentaries and response of the authors). It has led to the radical hypothesis that associative learning is of propositional nature, in other words that conditioning effects, at least in humans, actually emerge from propositional reasoning on beliefs about the causal nature of the environment [257]. This hypothesis is based on the assumption that associative learning is not automatic and requires awareness and effortful cognitive processes, though evidence regarding those claims is still unclear. However, our results favor a more modest view that learning does not systematically require the formation of low-level associations, but can also rely on high-level inference on abstract strategies. This hypothesis is consistent with multiple reports on the role of instructions and framing in experimental tasks (section 1.1), in addition to the ability to learn from self-generated hypotheses (section 1.2).

1.1 Instruction-driven biases

Humans notoriously exhibit framing biases in decision-making, in violation to normative economic theories [263]. For example, framing a lottery result in terms of losses or gains (for the same outcome) affects subjects' choices to gamble again [264]. Similarly, learning from positive or negative outcomes leads to different performances and

confidence judgements [265]. But framing effects exist beyond valence manipulations. Perceived value in a foraging context can vary according to the number of competitors [266] and the time horizon [267]. Instructions also alter decision making outside of value judgements. For example, Wyart and colleagues designed a task to study sequential processing of information when subjects were presented as actors generating the sequence or as passive observers [268, 269]. Interestingly, when observations are framed as actions' outcomes, participants tended to be more conservative against disconfirmatory evidence. This effect was likely explained by lower subjective volatility estimates when subjects perceived themselves as actors rather than passive observers [268].

Instructions can directly affect perceived volatility and feedback reliability in reversal learning tasks [270, 271]. But they can also alter the perception of the task structure itself. Model-free/model-based balance in reinforcement learning is frequently assessed using a simplified navigation task, the "*two-step task*" [207]. Although initial results demonstrated a mixture of model-free/model-based strategies, recent studies investigated the role of instructions and framing to manipulate the reliance on model-based learning [272, 273]. Feher da Silva & Hare successfully enhanced model-based behavior using elaborate cover stories that highlight the latent probabilistic connections between states in the task [272]. Moreover, they gave an alternative explanation to previous reports by simulating model-based agents with incorrect models of the task that reproduce patterns usually interpreted as hybrid model-free/model-based strategies. In contrast, Castro-Rodrigues and colleagues reported that participants with minimal instructions behave spontaneously as pure model-free reinforcement learners, and that detailed instructions provided latter could increase model-based behavior when the transition probabilities were stable enough [273]. They concluded, in opposition with Feher da Silva & Hare, that model-free reinforcement learning is the default strategy for solving the two-step task. However, their conclusions might be distorted by the lack of a thorough control of the pre-instructions experience, as instructions were always given after three uninstructed sessions. Indeed, the two-step task is highly simplified and barely resembles a navigation task, thus model-based planning rarely pays off compared to model-free learning [274]. Hence, reliance on a pure model-free strategy in the absence of instructions, might be due to low expected gains from model-based planning rather than a default strategy. This is supported by the finding that instructions enhanced model-based planning in this study only when the transition probabilities were stable, suggesting that, in the volatile condition, subjects actively discarded instructions due to already good results with a model-free strategy.

In our study, subjects were presented the same instructions, stimuli and feedbacks, in environments that varied solely in their latent statistical regularities. Yet, we found that they relied on global inferences rather than local associative links and that global effects were prone to strong within-subject variations, reactive to environmental differences. Interestingly, these variations were associated with variability in the weight of evidence and the reliability threshold of the strategies. This suggests that our participants were capable of generating and maintaining a space of hypotheses regarding the environment from which they could infer relevant strategies.

1.2 Self-generated hypotheses

Substantial evidence support the idea that associative learning relies on self-generated hypotheses. For example, in an instructed categorization task, some subjects discovered additional predictive information from alternative features of the stimuli, and used it to modify their behavior independently of explicit cues [242]. Moreover, constraints on the space of relevant features are necessary for the tractability of associative learning [127, 128]. Several reports suggest that subjects use bottom-up signals, such as prediction errors or expected value along individual dimensions, in order to reduce the space of relevant combinations [127, 259]. However, dimensionality reduction is a necessary condition for the generalizability of associative learning in a large state space and thus cannot rely only on the by-products of associative learning itself.

In a probabilistic learning task where stimulus-action mappings were based on two out of three dimensions of the stimuli, Cortese and colleagues showed that participants used abstract representations of the combinatorial space, rather than local feature learning [259]. They proposed a model initialized with a mixture of possible state representations, weighted by their reliability to predict upcoming outcomes (as in [201]). This is very similar to our approach. Indeed, even though our results strongly supports a representation at the policy level (the whole set of state-action mappings), it remains possible that subjects shortly envisioned both policy-based and value-based action space before committing fully to the former.

While policy-based inference does not preclude associative value learning, it can specifically account for several empirical findings, such as memory effects [247], counterfactual inference [113, 240] and range adaptation [117]. Most of mixture of experts models, based on inferences at the policy level, allow said policies to be adjusted locally [201, 249]. In our task, as in [259], the small number of possible stimulus-action mappings allowed a tractable representation of the complete combinatorial policy space, thus reducing the interest of local updates. Hence, in our paradigm, strategic inference merges with policy inference, though — in general — a representational frame might include strategies at different levels. Indeed, the functional organization of the prefrontal cortex might elicit nested representations, from elementary action selection to the selection of task-sets [244].

The question remains as to how humans are capable of building such abstract spaces. In the next section, we discuss recent progress in understanding the ability to manipulate abstract representations, with the aim of generating adapted strategies.

2 How to build a representational frame ?

For strategic inference to be truly adaptive, strategies must be constructed from abstract and flexible ingredients within the representational frame. First, strategies respond to specific **goals** by attributing **subjective value** to internal and external states (section 2.1). Second, strategies require **relational knowledge** about abstract objects, either to navigate between states or to create new objects from the basic properties of known

ones (section 2.2). Finally, building strategies, as well as hypotheses about the causal structure of the environment, is necessarily bounded by **computational constraints** (section 2.3).

2.1 Values and sub-goals generation

Value is a central component of learning, and adaptability is usually formalized as a reward maximization process. However, reward is also an elusive concept and its definition can be paradoxical[275]. Such contradictions emerge from the fact that learning has been conceived as a heterostatic process, aimed at maximising extrinsic values. As discussed in the introduction, this view has been explicitly advocated by Harry Klopf in the 1970s [19], and is compatible with the interpretation of rewards as hardwired reinforcing states. For example, perceiving the satisfaction of vital needs such as hunger or thirst as pleasant states could be the result of natural selection. At the same time, in natural environments, internal states play a crucial role in this valuation process : hunger, preferences, emotional ties, or political beliefs are not external factors and most of the rewarding stimuli in our everyday lives do not have a dedicated neural processing. Hence rewards must have intrinsic value, leading to a paradox [275].

Conversely, several authors have argued that learning emerges from the conservation of homeostasis. For example the free-energy principle [153] states that living organisms thrive to maintain a low surprise on internal and external states. Interestingly, under this hypothesis, learning can be reframed as active inference : actions are sampled in order to minimize the prediction error, thus policies are inferred as latent states [154]. Alternatively, Juechems & Summerfield recently proposed to extend the reinforcement learning theory to homeostatic constraints, following previous proposals [276, 275]. They note that : 1/ in natural environments, rewards cannot be given an extrinsic value only and need to refer to internal states and 2/ defining goals as *cognitive setpoints*, *i.e.* internally predefined objectives, reduces the computational complexity of learning a policy in high-dimensional environments. In this framework, setpoints are projected onto a multi-dimensional map, and the value of available options corresponds to their distance to a setpoint. This fits nicely with recent proposals that the orbitofrontal cortex, a region often involved in value-based decision making and reversal learning [277, 253], might represent a cognitive map of latent states [59, 60].

Indeed, the view that the OFC simply represents values is increasingly being questioned [278, 262]. For example, some authors have suggested that this region is used for goal-oriented policy selection, rather than option evaluation and value comparison [278, 279]. Functional neuro-imaging in humans showed that the patterns of activity in the OFC are compatible with the representation of hidden task-related states, rather than values *per se* [60]. In non human primates, neural activity preceding a choice in the medial prefrontal cortex have been shown to follow a ventro-dorsal gradient [280], consistent with the view of the most ventral part as a distant pre-motor structure [279]. This function of the OFC might be shared beyond primates. In rodents, the OFC has been linked with inferring value, rather than accessing previously cached values [58], and neural activity in the OFC has been shown to represent the possible destinations

available to the animal during a navigation task [281].

In machine learning, the interest for generating basis value functions for reinforcement learning has led to the formalization of *proto-value functions* [282]. This algorithm is based on the spectral decomposition of the transition matrix of the environment, *i.e.* the map, which allows to identify structural regularities and bottlenecks in the environment. The components of this decomposition can then serve as value basis functions, or proto-value functions, that corresponds to sub-goals in a large space. Interestingly, an analogous projection of the state space has recently been related to the activity of grid cells in the entorhinal cortex [283], which led to new hypotheses connecting value representation and cognitive maps [283, 284].

2.2 Conceptual maps and compositionality

It has long been suggested that the hippocampus encodes cognitive maps beyond the spatial domain [285, 286, 287, 288]. Maps or graphs are a natural formalism to model relationships between discrete entities such as locations, concepts or memories [288, 289, 284]. Hippocampus place cells have been shown to reactivate during sleep, hence replaying previous trajectories and simulating alternative options [290, 291]. Interestingly, this allows off-line learning, *i.e.* learning contingencies outside of their natural environment. Moreover, the hippocampus is sensitive to relational information beyond transition probabilities between states [292], and stores information in a format that supports generalization [293]. Taken together, these results led to the hypothesis that the hippocampus generates relational maps in novel contexts from the projection of representations about the structural properties of the environment in the OFC and entorhinal cortex [284, 294].

Alternatively, growing evidence points towards a fronto-parietal network for **compositional generalization** [295, 296]. Compositionality refers to the ability to partition abstract knowledge about previously encountered tasks, in order to recombine resulting subparts for generalization in novel environments [297, 298]. This is one of the most pressing topics in machine learning, since it is a skill in which humans still outperform artificial intelligence [298, 299]. Neural correlates of one-dimensional projections of ordinal concepts, such as values or numbers, have been found in the parietal cortex [300] in a format that keeps track of context-relevant information [301], and the patterns of activations in a fronto-parietal network are consistent with task-dependent low-dimensional projections [130]. In other words, the fronto-parietal cortex seems to encode rich information in an abstract format allowing for task-specificity and generalization [296].

Recent efforts aimed at unifying reinforcement learning and abstract knowledge about the structure of the environment under the terminology of *theory-based reinforcement learning* [302]. This approach extends model-based reinforcement learning by providing the agent with a specific object-oriented theory, using the Video Game Description Language [303]. Analysis of fMRI data while playing Atari-style video games revealed theory-related activity in a large fronto-parietal network, but not the hippocampus [304]. This could suggest a functional distinction between conceptual maps involving the entorhinal-hippocampal (EC-HC) network and compositionality of structural knowledge related to

a parieto-prefrontal network. The relative implication of both networks for generalization might be attributed to divergent evolutionary paths. As mentioned in [296], rodents mostly interact with the world via locomotion, promoting an allocentric representation involving the EC-HC network, while primates rely more on manipulations in the surrounding (abstract) space, likely supported by the parieto-prefrontal network [295, 296].

2.3 Considerations on the computational complexity

Computational complexity refers to the overall evaluation of resources required for a given problem. Despite the extensive use of computational modeling in cognitive science, formal assessment of tractability is often overlooked [305]. Classically, problems have different arguments, or inputs, each of various sizes. A tractable problem usually requires resources that do not grow too much given the size of its arguments. More formally, the maximal running time of a polynomial-time algorithm follows a polynomial function of its input size. Such problems are considered tractable. Other problems, like finding the best explanation in a set of hypotheses using Bayesian inference, have a running time that grows exponentially with the size of the input (*e.g.* the number of hypotheses, the dimensionality of the observations...) [155]. This does not mean that every instance of Bayesian inference is intractable, which would obviously be wrong, but that unbounded Bayesian inference, *i.e.* without constraints on the input size, is not tractable. Importantly, this result does not depend on the specific algorithms or approximations that are used [155, 236].

Besides this general approach, that often yields negative (*i.e.* intractability) results, complexity analysis also allows to explore the individual boundaries required for each argument. This is called *parameterized complexity analysis* and can provide *fixed-parameter tractability* results, as one tries to impose constraints on each parameter of the problem, in other words to characterize which instances are tractable and which are not [305]. Regarding Bayesian inference, fixed-parameter tractability results vary, according to algorithmic choices, but they all impose, among other constraints, a relatively small hypotheses space (for details see [155]).

Without providing a formal complexity analysis of our model, we can draw the attention on the specific computational cost of strategic inference, compared to associative learning. Indeed, the strategic space must be bounded in order for Bayesian inference to be tractable. Yet, in our model we used the full combinatorial space of potential stimulus-action mappings that is relatively low in our experimental set-up, but grows exponentially with the number of stimuli. It is then possible that increasing the number of states or the dimensionality of stimuli might lead strategic inference to completely collapse in favor of more local learning schemes. This is however unlikely since the curse of dimensionality affects associative learning as well [2]. Another possibility is that subjects do not explore the complete space of possible strategies, but only a part of this space. This would require an arbitrary partition of the strategic space between what is possible and what is not. In other words, in order for strategic inference to generalize to natural environments, it must rely on *guesses* of the underlying strategic space.

The formal complexity of forming ideas, guesses or generating a space of hypothesis

is of a completely different order of magnitude than inference to the best explanation [306]. Yet, it constitutes a central part of our abilities to communicate or understand the world. A critical consequence of recasting meta-learning as strategic inference, is that it makes it dependent of such processes.

3 Perspectives for future work

3.1 Behavioral and computational investigations

In this work, we captured within-subjects variations in response to covert contextual factors. Importantly, we did not change the instructions nor the stimuli shown to our participants. However, as we discussed earlier, framing effects are pervasive in decision-making tasks, and future work might take advantage of this to further our understanding of top-down processes in meta-learning. Indeed, the degree of flexibility of the representational frame remains an open question. In particular, as in most (but not all) studies in the reinforcement learning field, we focused on discrete actions with binary feedbacks. This gave clear contours for constructing a finite space of discrete strategies. In a setup with continuous state-spaces or more complex feedbacks, strategic inference might rely on more arbitrary and idiosyncratic delineations of the strategic repertoire.

This issue is directly related to the computational constraints on the representation frame. In our model, we have simply instantiated the full space of possible strategies, although this is not possible for large environments, nor is it necessary. Fundamentally, strategic inference relies on initial *guesses* over the prior space of possibilities. Computational accounts of guessing and abduction (*i.e.* hypotheses generation) are still in their infancy [299, 306]. However, one possibility to explore the strategic repertoire generated by human subjects could be to use a descriptive Bayesian modelling of choices (*e.g.* a hidden Markov model) as we did in our study, for large but finite possible spaces. Interestingly, recent proposals have put forward such state-space models in animal literature [307, 308].

Moreover, the nature of competing strategies might differ, *e.g.* alternating between model-free and model-based control or coordinating action chuncks of various depths [309]. Importantly, qualitatively different strategies make different predictions regarding cost/benefit trade-offs, that could be tested experimentally. For example, while model-free reinforcement learning predicts constant computational cost for updating local associations, model-based planning requires a variable quantity of resources for inferring prospective values [214]. In between, strategic inference over pre-specified policies predicts a constant cost for strategy monitoring, and occasional overhead for strategic construction or for reconfigurations of the representational frame. These distinctive signatures could give interesting predictions for future experiments.

3.2 Neural correlates

Our proposal that meta-learning is based on a flexible representational frame is consistent with other suggestions about the versatility of the prefrontal cortex for de-

cision making (*e.g.* [262, 279]). However, these ideas are often supported by the accumulation of experimental protocols focused on specific questions, rather than the direct demonstration of context-dependent reconfiguration of neural processes in the same experiment. A unified account would predict that regions usually involved in economic decision, counterfactual inference and set-shifting, are part of a network dedicated to strategic inference. One would expect not only correlations between internal variables of the model and neural activity in these areas, but also the persistence of these correlations in various contexts.

In addition, the strategic inference model makes valuable predictions regarding the role of the neurotransmitters such as dopamine and norepinephrine. Indeed, dopamine has been associated with update signals in associative learning, but also with state-based inference [166]. Interestingly, strategic inference merges these processes, as it optimizes action selection by inferring the latent rule. Our model also produces random exploration as a direct response to ambiguity (*i.e.* when observations cannot improve the distinction between candidate strategies). This is consistent with the many roles attributed to noradrenergic inputs to the dmPFC, either as reset signals [77, 104], uncertainty monitoring [75] or internal learning noise [100].

While imprecision in low-level updates has been suggested as a potential confound for high-level volatility estimates for exact strategic inference [145], our results show that long-term contextual statistics affect human behavior despite similar short-term conditions. This is not compatible with emergent adaptation from local update noise only, since, locally, both situations were undistinguishable and would have generated similar inference noise. However, since our model is top-down and does not rely on associative learning, it is compatible with updating noise as a general feature, independently from the representational level of inference (local associations or global policies). A natural extension of the model would then be to add random noise to the update of strategic reliabilities and assess its correlation with noradrenergic activity. Crucially, this correlation should not depend on the scope of the problem at hand (local or global), and help clarify the relative contribution of external and internal noise to strategic ambiguity.

3.3 Evolutionary and translational perspectives

Translational research is based on functional homologies between brain regions. The use of animal models to study human cognition is therefore limited by assumptions about the conservation of similar functions across divergent evolutionary lineages. Interestingly, several marker of strategic inference have been found in non human primates. In a task where macaque monkeys had to learn successive rules that were initially cued and then switched without notice, monkeys showed increased trap reactivity during the initial phase, suggesting an anticipatory adaptation to unpredictible switches [241]. Other studies have highlighted the use of abstract context representations [196, 240], fast behavioral switches [190] and information seeking [310] in monkeys. These phenomena require the arbitration between abstract strategies, likely supported by the activity of medial prefrontal areas [190, 191, 196, 240, 310].

While humans and non human primates share a highly homologous organization of

their prefrontal cortex [199, 311, 295], such similarities between rodents and primates are still debated [312, 313, 314]. Indeed, goal generation and fast generalizations have been suggested to be specific to primates, via the development of a parieto-prefrontal network [295]. Hence, two possibilities can be formulated regarding strategic inference in rodents. Either they rely more on associative learning and less on strategic inference, or they use strategic inference extensively, but on a radically different repertoire of strategies than primates do. We lean towards the latter, for three main reasons.

First, rodents have been shown to relate on homologous cortical and subcortical regions than humans for solving complex tasks [315, 316]. For example, in rats, medial prefrontal regions are causally involved in inferring values [58] and shifting from one strategy to the other [80]. Second, rodents can come up with original heuristics when solving complex tasks. Latent state modelling of mice behavior revealed that, during the course of a perceptual decision making task, mice alternate between different strategies [308]. Moreover, in a experimental paradigm similar to ours, mice displayed specific strategies for learning stimulus-action contingencies after partial rule changes [317] : while they did not show interference effects as our subjects, they learned local rules with variable delays, depending on the local association rules. Indeed, actions that were correct for only one stimulus were learned considerably slower than actions that were common to other stimuli. This suggests that mice reduce the observable state space by reasoning in terms of local strategies, rather than independent stimulus-action associations. Finally, flexible behavior and the acquisition of complex strategies exist in more phylogenetically distant species such as birds. Birds are capable of using tools [318], and display flexible behavior in natural and laboratory contexts [319, 320].

Therefore, extending the model of strategic inference beyond human cognition would involve both identifying computational functions that have been conserved over the course of evolution and exploring the strategic repertoires specific to each species, depending on the constraints imposed by its natural environment.

Conclusion

I introduced the topic of meta-learning with the example of a doctoral student and her supervisor. Both of them can learn, and learn to learn, by incremental trials and errors, starting with minimal assumptions and expectations. This idea of building scientific knowledge (about learning) directly from the data, has been advocated by Brutus Skinner in his 1950 paper "*Are theories of learning necessary ?*" [321]. He warned *against* the use of theories to orient data collection, while acknowledging the value of formal analysis and intuitive exploration of the data. This seems an incongruous warning today, when many voices are being raised to denounce the crisis of replication in psychology [322], and to demonstrate the importance of multiplying theoretical viewpoints for advancing cognitive science [323, 324]. Contrary to Skinner's view, learning theories are certainly not just fun but undoubtedly necessary. They are also probably unavoidable, as no single empirical fact can provide a unique interpretation by itself.

When interpreting the results of his experiments, Edward Thorndike was struck by the inability of the animals to make connections between the various situations they were in. Unlike a human being, who is able to understand his or her environment, to have ideas, Thorndike's subjects seemed to be sensitive only to the repetition of stereotyped situations. While sharing his views, Margaret Washburn was more cautious than Thorndide. She noted that, in natural environments, animals were able to adapt quickly, without intensive repetition, and that being unable to understand a man-made puzzle box was perhaps not enough to conclude that they were incapable of having ideas. But the early debates in comparative psychology about the existence of abstract ideas in animals are probably ill-posed and symptomatic of an anthropocentric interpretation of evolution.

While the PhD student and her supervisor will gradually improve their ability to make sense of their data, they did not start from nowhere. Shared knowledge, social constraints and personal preferences initially influenced their experimental design and will affect their perception of their results. Similarly, animals do not adapt to a new environment by applying generic algorithms from a blank state. In this work, we showed that strategic inference better captures human adaptability in complex tasks than incremental associative learning. Moreover, strategic inference itself proved to be very adaptable to covert environmental regularities. We suggested that accounting for this flexibility requires the use of abstract knowledge about the world, following many other

authors in the field of natural and artificial intelligence.

The term "*representational frame*" aims at covering these overarching cognitive processes which converge towards a finite strategic repertoire. This can be interpreted as having ideas or theories about the world, though it can also include more basic components. Thus, although learning in the simplified setting of a laboratory task may involve mechanisms that are highly conserved during evolution, different species may have evolved very different approaches to meta-learning via the construction of divergent representational frames.

It could be argued that this gives too much explanatory weight to the representational frame, and only postpones the problem of meta-learning. However, several non-trivial questions remain about strategic inference — its interaction with associative learning and its neural implementation — which do not require engaging with the representational frame. Nevertheless, the emergence of symbolic representations is indeed the most difficult question in cognitive science and, at the same time, the most indispensable ingredient for its advancement [325].

Bibliographie

[1] Clark L Hull. *Principles of behavior*. 1943.

[2] Richard S Sutton and Andrew G Barto. *Reinforcement learning : An introduction*. MIT press, 2018.

[3] Marvin Minsky. Steps Toward Artificial Intelligence. *Proceedings of the IRE*, 49(1) :8–30, 1961.

[4] W. Schultz, P. Dayan, and P. R. Montague. A neural substrate of prediction and reward. *Science*, 275(5306) :1593–1599, 1997.

[5] Edward L Thorndike. *Animal Intelligence : experimental studies*. New York, 1911.

[6] Ann G. Winfield. *Resuscitating bad science : Eugenics past and present*. Teachers College Press, New York, w.h., edit edition, 2012.

[7] Margaret Washburn. *The Animal Mind*. 1923.

[8] R. H. Waters. The law of effect as a principle of learning. *Psychological Bulletin*, 31(6) :408–425, 1934.

[9] Ivan Petrovitch Pavlov and William Gantt. Lectures on conditioned reflexes : Twenty-five years of objective study of the higher nervous activity (behaviour) of animals. 1928.

[10] Burrhus F Skinner. Two types of conditioned reflex : A reply to Konorski and Miller. *The Journal of General Psychology*, 16(1) :272–279, 1937.

[11] Robert Bush and Frederick Mosteller. *Stochastic models for learning*. 1955.

[12] Robert A Rescorla and Allan R Wagner. A theory of Pavlovian conditioning : The effectiveness of reinforcement and non- reinforcement. In *Classical Conditioning II : Current Theory and Research*, number January 1972. 1972.

[13] Leon J Kamin. Predictability, surprise, attention, and conditioning. In *SYMP. ON PUNISHMENT*, number TR-13, 1967.

[14] Leon J Kamin. Selective association and conditioning. *Fundamental issues in associative learning*, pages 42–64, 1969.

[15] Richard Bellman. A Markovian Decision Process. *Journal of Mathematics and Mechanics*, pages 679–684, 1957.

[16] Stuart Dreyfus. Richard Bellman on the birth of dynamic programming. *Operations Research*, 50(1) :48–51, 2002.

[17] Ian H. Witten. An adaptive optimal controller for discrete-time Markov environments. *Information and Control*, 34(4) :286–295, 1977.

[18] Richard S. Sutton. Learning to Predict by the Methods of Temporal Differences. *Machine Learning*, 3(1) :9–44, 1988.

[19] A Harry Klopf. *Brain function and adaptive systems : a heterostatic theory*. Number 133. Air Force Cambridge Research Laboratories, Air Force Systems Command, United . . . , 1972.

[20] Christopher J C H Watkins. Learning from delayed rewards. PhD thesis, 1989.

[21] Christopher J C H Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3) :279–292, 1992.

[22] R Duncan Luce. *Individual choice behavior : A theoretical analysis*. Courier Corporation, 2012.

[23] Daniel L McFadden. Quantal choice analaysis : A survey. *Annals of Economic and Social Measurement, Volume 5, number 4*, pages 363–390, 1976.

[24] Ranulfo Romo and Wolfram Schultz. Dopamine neurons of the monkey midbrain : Contingencies of responses to stimuli eliciting immediate behavioral reactions. *Journal of Neurophysiology*, 63(3) :607–624, 1990.

[25] T. Ljungberg, P. Apicella, and W. Schultz. Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of Neurophysiology*, 67(1) :145–163, 1992.

[26] J. Mirenowicz and W. Schultz. Importance of unpredictability for reward responses in primate dopamine neurons. *Journal of Neurophysiology*, 72(2) :1024–1027, 1994.

[27] Wolfram Schultz, Léon Tremblay, and Jeffrey R. Hollerman. Reward prediction in primate basal ganglia and frontal cortex. *Neuropharmacology*, 37(4-5) :421–429, 1998.

[28] Gordon J. Mogenson, Douglas L. Jones, and Chi Yiu Yim. From motivation to action : Functional interface between the limbic system and the motor system. *Progress in Neurobiology*, 14(2-3) :69–97, 1980.

[29] Matthijs A.A. Van der Meer and A. David Redish. Ventral striatum : A critical look at models of learning and evaluation, 2011.

[30] Andrew G Barto. Adaptive Critics and the Basal Ganglia. *Models of information processing in the basal ganglia*, 215, 1995.

[31] Daphna Joel, Yael Niv, and Eytan Ruppin. Actor-critic models of the basal ganglia : new anatomical and computational perspectives. *Neural Networks*, 15(4-6) :535–547, 2002.

[32] John O'Doherty, Peter Dayan, Johannes Schultz, Ralf Deichmann, Karl Friston, and Raymond J. Dolan. Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning. *Science*, 304(5669) :452–454, 2004.

[33] Mehdi Khamassi, Loic Lacheze, Benoit Girard, Alain Berthoz, and Agnes Guillot. Actor-critic models of reinforcement learning in the basal ganglia : From natural to artificial rats. *Adaptive Behavior*, 13(2) :131–148, 2005.

[34] Joseph F. Cheer, Brandon J. Aragona, Michael L.A.V. Heien, Andrew T. Seipel, Regina M. Carelli, and R. Mark Wightman. Coordinated Accumbal Dopamine Release and Neural Activity Drive Goal-Directed Behavior. *Neuron*, 54(2) :237–244, 2007.

[35] Jeremy J. Day, Mitchell F. Roitman, R. Mark Wightman, and Regina M. Carelli. Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. *Nature Neuroscience*, 10(8) :1020–1028, 2007.

[36] Giuseppe Pagnoni, Caroline F. Zink, P. Read Montague, and Gregory S. Berns. Activity in human ventral striatum locked to errors of reward prediction. *Nature Neuroscience*, 5(2) :97–98, 2002.

[37] Mathias Pessiglione, Ben Seymour, Guillaume Flandin, Raymond J. Dolan, and Chris D. Frith. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, 442(7106) :1042–1045, 2006.

[38] Todd A. Hare, John O'Doherty, Colin F. Camerer, Wolfram Schultz, and Antonio Rangel. Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *Journal of Neuroscience*, 28(22) :5623–5630, 2008.

[39] Mehdi Khamassi, Antonius B Mulder, Eiichi Tabuchi, Vincent Douchamps, and Sidney I Wiener. Anticipatory reward signals in ventral striatal neurons of behaving rats. *European Journal of Neuroscience*, 28(9) :1849–1866, 2008.

[40] Hannah M. Bayer and Paul W. Glimcher. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, 47(1) :129–141, 2005.

[41] Juliana Yacubian, Jan Gläscher, Katrin Schroeder, Tobias Sommer, Dieter F Braus, and Christian Büchel. Dissociable systems for gain- and loss-related value predictions and errors of prediction in the human brain. *Journal of Neuroscience*, 26(37) :9530–9537, 2006.

[42] Nathaniel D. Daw, Sham Kakade, and Peter Dayan. Opponent interactions between serotonin and dopamine. *Neural Networks*, 15(4-6) :603–616, 2002.

[43] Stefano Palminteri, Anne Helene Clair, Luc Mallet, and Mathias Pessiglione. Similar improvement of reward and punishment learning by serotonin reuptake inhibitors in obsessive-compulsive disorder. *Biological Psychiatry*, 72(3) :244–250, 2012.

[44] Stefano Palminteri, Damian Justo, Celine Jauffret, Beth Pavlicek, Aurelie Dauta, Christine Delmaire, Virginie Czernecki, Carine Karachi, Laurent Capelle, Alexandra Durr, and Mathias Pessiglione. Critical Roles for Anterior Insula and Dorsal Striatum in Punishment-Based Avoidance Learning. *Neuron*, 76(5) :998–1009, 2012.

[45] Maelle C.M. Gueguen, Alizée Lopez-Persem, Pablo Billeke, Jean Philippe Lachaux, Sylvain Rheims, Philippe Kahane, Lorella Minotti, Olivier David, Mathias Pessiglione, and Julien Bastin. Anatomical dissociation of intracerebral signals for reward and punishment prediction errors in humans. *Nature Communications*, 12(1) :1–12, 2021.

[46] Darrell Haufler, Omer Liran, Robert J Buchanan, and Denis Pare. Human anterior insula signals salience and deviations from expectations via bursts of beta oscillations. *Journal of Neurophysiology*, 128(1) :160–180, 2022.

[47] Kerstin Preuschoff, Steven R Quartz, and Peter Bossaerts. Human insula activation reflects risk prediction errors as well as risk. *Journal of Neuroscience*, 28(11) :2745–2752, 2008.

[48] Peter Bossaerts. Risk and risk prediction error signals in anterior insula., 2010.

[49] Masayuki Matsumoto and Okihide Hikosaka. Lateral habenula as a source of negative reward signals in dopamine neurons. *Nature*, 447(7148) :1111–1115, 2007.

[50] Ethan S. Bromberg-Martin, Masayuki Matsumoto, and Okihide Hikosaka. Dopamine in Motivational Control : Rewarding, Aversive, and Alerting. *Neuron*, 68(5) :815–834, 2010.

[51] Okihide Hikosaka. The habenula : from stress evasion to value-based decision-making. *Nature reviews neuroscience*, 11(7) :503–513, 2010.

[52] Stephan Lammel, Byung Kook Lim, Chen Ran, Kee Wui Huang, Michael J. Betley, Kay M. Tye, Karl Deisseroth, and Robert C. Malenka. Input-specific control of reward and aversion in the ventral tegmental area. *Nature*, 491(7423) :212–217, 2012.

[53] Hackjin Kim, Shinsuke Shimojo, and John P. O'Doherty. Overlapping responses for the expectation of juice and money rewards in human ventromedial prefrontal cortex. *Cerebral Cortex*, 21(4) :769–776, 2011.

[54] Dino J Levy and Paul W Glimcher. The root of all value : a neural common currency for choice. *Current opinion in neurobiology*, 22(6) :1027–1038, 2012.

[55] Daniel McNamee, Antonio Rangel, and John P O'doherty. Category-dependent and category-independent goal-value codes in human ventromedial prefrontal cortex. *Nature neuroscience*, 16(4) :479–485, 2013.

[56] Jan Gläscher, Ralph Adolphs, Hanna Damasio, Antoine Bechara, David Rudrauf, Matthew Calamia, Lynn K. Paul, and Daniel Tranel. Lesion mapping of cognitive control and value-based decision making in the prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 109(36) :14681–14686, 2012.

[57] M. P. Noonan, M. E. Walton, T. E.J. Behrens, J. Sallet, M. J. Buckley, and M. F.S. Rushworth. Separate value comparison and learning mechanisms in macaque medial and lateral orbitofrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 107(47) :20547–20552, 2010.

[58] Joshua L Jones, Guillem R Esber, Michael A McDannald, Aaron J Gruber, Alex Hernandez, Aaron Mirenzi, and Geoffrey Schoenbaum. Orbitofrontal cortex supports behavior and learning using inferred but not cached values. *Science*, 338(6109) :953–956, 2012.

[59] Robert C. Wilson, Yuji K. Takahashi, Geoffrey Schoenbaum, and Yael Niv. Orbitofrontal cortex as a cognitive map of task space. *Neuron*, 81(2) :267–279, 2014.

[60] Nicolas W. Schuck, Ming Bo Cai, Robert C. Wilson, and Yael Niv. Human Orbitofrontal Cortex Represents a Cognitive Map of State Space. *Neuron*, 91(6) :1402–1412, 2016.

[61] Philippe Domenech and Etienne Koechlin. Executive control and decision-making in the prefrontal cortex. *Current Opinion in Behavioral Sciences*, 1 :101–106, 2015.

[62] John M. Pearce and Geoffrey Hall. A model for Pavlovian learning : Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, 87(6) :532–552, 1980.

[63] John M Pearce, Helen Kaye, and Geoffrey Hall. Predictive accuracy and stimulus associability : Development of a model for Pavlovian learning. *Quantitative analyses of behavior*, 3 :241–256, 1982.

[64] Matthew R Roesch, Guillem R Esber, Jian Li, Nathaniel D Daw, and Geoffrey Schoenbaum. Surprise ! neural correlates of pearce–hall and rescorla–wagner coexist within the brain. *European Journal of Neuroscience*, 35(7) :1190–1200, 2012.

[65] Richard S. Sutton. Adapting bias by gradient descent : an incremental version of delta-bar-delta. *Proceedings Tenth National Conference on Artificial Intelligence*, pages 171–176, 1992.

[66] Noboru Murata, Motoaki Kawanabe, Andreas Ziehe, Klaus-Robert Müller, and Shun-ichi Amari. On-line learning in changing environments with applications in supervised and unsupervised learning. *Neural Networks*, 15(4-6) :743–760, 2002.

[67] Kenji Doya. Metalearning and neuromodulation. *Neural Networks*, 15 :495–506, 2002.

[68] Toshihiko Aosaki, Hiroshi Tsubokawa, Akihiro Ishida, Katsushige Watanabe, Ann M. Graybiel, and Minoru Kimura. Responses of tonically active neurons in the primate's striatum undergo systematic changes during behavioral sensorimotor conditioning. *Journal of Neuroscience*, 14(6) :3969–3984, 1994.

[69] John G. Partridge, Subbu Apparsundaram, Greg A. Gerhardt, Jennifer Ronesi, and David M. Lovinger. Nicotinic Acetylcholine Receptors Interact with Dopamine in Induction of Striatal Long-Term Depression. *Journal of Neuroscience*, 22(7) :2541–2549, 2002.

[70] Angela Yu and Peter Dayan. Expected and unexpected uncertainty : ACh and NE in the neocortex. In *Advances in Neural Information Processing Systems*, 2003.

[71] Angela J. Yu and Peter Dayan. Uncertainty, neuromodulation, and attention. *Neuron*, 46(4) :681–692, 2005.

[72] Nicolas Schweighofer, Saori C Tanaka, and Kenji Doya. Serotonin and the evaluation of future rewards : theory, experiments, and possible neural mechanisms. *Annals of the New York Academy of Sciences*, 1104(1) :289–300, 2007.

[73] Hanneke E.M. DenOuden, Nathaniel D. Daw, Guillén Fernandez, Joris A. Elshout, Mark Rijpkema, Martine Hoogman, Barbara Franke, and Roshan Cools. Dissociable Effects of Dopamine and Serotonin on Reversal Learning. *Neuron*, 80(4) :1090–1100, 2013.

[74] Benjamin Y. Hayden, Sarah R. Heilbronner, John M. Pearson, and Michael L. Platt. Surprise signals in anterior cingulate cortex : Neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. *Journal of Neuroscience*, 31(11) :4178–4187, 2011.

[75] Massimo Silvetti, Ruth Seurinck, Marlies E. van Bochove, and Tom Verguts. The influence of the noradrenergic system on optimal control of neural plasticity. *Frontiers in Behavioral Neuroscience*, 7(NOV) :1–6, 2013.

[76] Marieke Jepma, Peter R Murphy, Matthew R Nassar, Mauricio Rangel-Gomez, Martijn Meeter, and Sander Nieuwenhuis. Catecholaminergic Regulation of Learning Rate in a Dynamic Environment. *PLoS Computational Biology*, 12(10) :1–24, 2016.

[77] Sebastien Bouret and Susan J. Sara. Network reset : A simplified overarching theory of locus coeruleus noradrenaline function. *Trends in Neurosciences*, 28(11) :574–582, 2005.

[78] Massimo Silvetti, Eliana Vassena, Elger Abrahamse, and Tom Verguts. Dorsal anterior cingulate-midbrain ensemble as a reinforcement meta-learner. *PLoS computational biology,*, 14(8) :e1006370, 2018.

[79] Gina R Poe, Stephen Foote, Oxana Eschenko, Joshua P Johansen, Sebastien Bouret, Gary Aston-Jones, Carolyn W Harley, Denise Manahan-Vaughan, David Weinshenker, Rita Valentino, Craig Berridge, Daniel J Chandler, Barry Waterhouse, and Susan J. Sara. Locus coeruleus : a new look at the blue spot. *Nature Reviews Neuroscience*, 21(11) :644–659, 2020.

[80] D Gowanlock R Tervo, Elena Kuleshova, Maxim Manakov, Mikhail Proskurin, Mattias Karlsson, Andy Lustig, Reza Behnam, and Alla Y Karpova. The anterior cingulate cortex directs exploration of alternative strategies. *Neuron*, 109(11) :1876–1887.e6, 2021.

[81] Steven W. Kennerley, Mark E. Walton, Timothy E.J. Behrens, Mark J. Buckley, and Matthew F.S. Rushworth. Optimal decision making and the anterior cingulate cortex. *Nature Neuroscience*, 9(7) :940–947, 2006.

[82] Timothy E.J. Behrens, Mark W. Woolrich, Mark E. Walton, and Matthew F.S. Rushworth. Learning the value of information in an uncertain world. *Nature Neuroscience*, 10(9) :1214–1221, 2007.

[83] Marco K. Wittmann, Nils Kolling, Rei Akaishi, Bolton K.H. Chau, Joshua W. Brown, Natalie Nelissen, and Matthew F.S. Rushworth. Predictive decision making

driven by multiple time-linked reward representations in the anterior cingulate cortex. *Nature Communications*, 7(1) :1–13, 2016.

[84] Alberto Bernacchia, Hyojung Seo, Daeyeol Lee, and Xiao Jing Wang. A reservoir of time constants for memory traces in cortical neurons. *Nature Neuroscience*, 14(3) :366–372, 2011.

[85] Adam S. Lowet, Qiao Zheng, Sara Matias, Jan Drugowitsch, and Naoshige Uchida. Distributional Reinforcement Learning in the Brain. *Trends in Neurosciences*, 43(12) :980–997, 2020.

[86] Shiva Farashahi, Christopher H Donahue, Peyman Khorsand, Hyojung Seo, Daeyeol Lee, and Alireza Soltani. Metaplasticity as a Neural Substrate for Adaptive Learning and Choice under Uncertainty. *Neuron*, 94(2) :401–414.e6, 2017.

[87] J.G. March. Exploration and exploitation in organizational learning. *Organization Science*, 2(1) :71 – 87, 1991.

[88] Sebastian B Thrun. Efficient Exploration In Reinforcement Learning. (January) :1–44, 1992.

[89] Nathaniel D Daw, John P. O'Doherty, Peter Dayan, Ben Seymour, and Raymond J Dolan. Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095) :876–879, 2006.

[90] Robert C WILSON, Andra GEANA, John M WHITE, Elliot A LUDVIG, and Jonathan D COHEN. Humans Use Directed and Random Exploration to Solve the Explore-Exploit Dilemma. *Journal of experimental psychology. General*, 143(6) :2074–2081, 2014.

[91] Samuel J Gershman. Deconstructing the human algorithms for exploration. *Cognition*, 173 :34–42, 2018.

[92] Robert C Wilson, Elizabeth Bonawitz, Vincent D Costa, and R. Becket Ebitz. Balancing exploration and exploitation with information and randomization, 2021.

[93] David Badre, Bradley B. Doll, Nicole M. Long, and Michael J. Frank. Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. *Neuron*, 73(3) :595–607, 2012.

[94] Momchil S. Tomov, Van Q. Truong, Rohan A. Hundia, and Samuel J. Gershman. Dissociable neural correlates of uncertainty underlie different exploration strategies. *Nature Communications*, 11(1), 2020.

[95] Mael Donoso, Anne G.E. Collins, and Etienne Koechlin. Foundations of human reasoning in the prefrontal cortex. *Science*, 344(6191) :1481–1486, 2014.

[96] Kenji Doya. Modulators of decision making. *Nature Neuroscience*, 11(4) :410–416, 2008.

[97] Marieke Jepma and Sander Nieuwenhuis. Pupil diameter predicts changes in the exploration-exploitation trade-off : Evidence for the adaptive gain theory. *Journal of Cognitive Neuroscience*, 23(7) :1587–1596, 2011.

[98] Marieke Jepma, Erik T. te Beek, Eric Jan Wagenmakers, Joop M.A. van Gerven, and Sander Nieuwenhuis. The role of the noradrenergic system in the exploration-exploitation trade-off : A psychopharmacological study. *Frontiers in Human Neuroscience*, 4(August) :1–13, 2010.

[99] Christopher M Warren, Robert C Wilson, Nic J. Van Der Wee, Eric J Giltay, Martijn S. Van Noorden, Jonathan D Cohen, and Sander Nieuwenhuis. The effect of atomoxetine on random and directed exploration in humans. *PLoS ONE*, 12(4) :1–17, 2017.

[100] Charles Findling, Vasilisa Skvortsova, Remi Dromnelle, Stefano Palminteri, and Valentin Wyart. Computational noise in reward-guided learning drives behavioral variability in volatile environments. *Nature Neuroscience*, 22(12) :2066–2077, 2019.

[101] Charles Findling and Valentin Wyart. Computation noise in human learning and decision-making : origin, impact, function. *Current Opinion in Behavioral Sciences*, 38 :124–132, 2021.

[102] Konstantinos Tsetsos, Rani Moran, James Moreland, Nick Chater, Marius Usher, and Christopher Summerfield. Economic irrationality is optimal during noisy decision making. *Proceedings of the National Academy of Sciences of the United States of America*, 113(11) :3102–3107, 2016.

[103] Jan Drugowitsch, Valentin Wyart, Anne Dominique Devauchelle, and Etienne Koechlin. Computational Precision of Mental Inference as Critical Source of Human Choice Suboptimality. *Neuron*, 92(6) :1398–1411, 2016.

[104] Dougal G.R. Tervo, Mikhail Proskurin, Maxim Manakov, Mayank Kabra, Alison Vollmer, Kristin Branson, and Alla Y Karpova. Behavioral variability through stochastic choice and its gating by anterior cingulate cortex. *Cell*, 159(1) :21–32, 2014.

[105] Nicolas Schweighofer and Kenji Doya. Meta-learning in reinforcement learning. *Neural Networks*, 16(1) :5–9, 2003.

[106] Mehdi Khamassi, Pierre Enel, Peter Ford Dominey, and Emmanuel Procyk. Medial prefrontal cortex and the adaptive regulation of reinforcement learning parameters. In *Progress in Brain Research*. 2013.

[107] Mehdi Khamassi, George Velentzas, Theodore Tsitsimis, and Costas Tzafestas. Active exploration and parameterized reinforcement learning applied to a simulated human-robot interaction task. In *Proceedings - 2017 1st IEEE International Conference on Robotic Computing, IRC 2017*, 2017.

[108] Mark D. Humphries, Mehdi Khamassi, and Kevin Gurney. Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. *Frontiers in Neuroscience*, (FEB), 2012.

[109] Christopher D. Fiorillo, Philippe N. Tobler, and Wolfram Schultz. Discrete Coding of Reward Dopamine Neurons. *Science*, 299(March) :1898–1902, 2003.

[110] Francois Cinotti, Virginie Fresno, Nassim Aklil, Etienne Coutureau, Benoit Girard, Alain R Marchand, and Mehdi Khamassi. Dopamine blockade impairs the exploration-exploitation trade-off in rats. *Scientific Reports*, 9(1) :1–14, 2019.

[111] Terry Lohrenz, Kevin McCabe, Colin F. Camerer, and P. Read Montague. Neural signature of fictive learning signals in a sequential investment task. *Proceedings of the National Academy of Sciences of the United States of America*, 104(22) :9493–9498, 2007.

[112] Erie D Boorman, Timothy E J Behrens, Mark W Woolrich, and Matthew F S Rushworth. Article How Green Is the Grass on the Other Side ? Frontopolar Cortex and the Evidence in Favor of Alternative Courses of Action. *Neuron*, 62(5) :733–743, 2009.

[113] Erie D. Boorman, Timothy E. Behrens, and Matthew F. Rushworth. Counterfactual choice and learning in a Neural Network centered on human lateral frontopolar cortex. *PLoS Biology*, 9(6), 2011.

[114] Adrian G. Fischer and Markus Ullsperger. Real and fictive outcomes are processed differently but converge on a common adaptive mechanism. *Neuron*, 79(6) :1243–1255, 2013.

[115] Stefano Palminteri, Germain Lefebvre, Emma J. Kilford, and Sarah Jayne Blakemore. Confirmation bias in human reinforcement learning : Evidence from counterfactual feedback processing. *PLoS computational biology*, 13(8) :e1005684, 2017.

[116] Antonio Rangel and John A. Clithero. Value normalization in decision making : Theory and evidence. *Current Opinion in Neurobiology*, 22(6) :970–981, 2012.

[117] Stefano Palminteri and Mael Lebreton. Context-dependent outcome encoding in human reinforcement learning. *Current Opinion in Behavioral Sciences*, 41 :144–151, 2021.

[118] Stefano Palminteri, Mehdi Khamassi, Mateus Joffily, and Giorgio Coricelli. Contextual modulation of value signals in reward and punishment learning. *Nature Communications*, 6 :1–14, 2015.

[119] Tilmann A. Klein, Markus Ullsperger, and Gerhard Jocham. Learning relative values in the striatum induces violations of normative decision making. *Nature Communications*, 8, 2017.

[120] Sophie Bavard, Mael Lebreton, Mehdi Khamassi, Giorgio Coricelli, and Stefano Palminteri. Reference-point centering and range-adaptation enhance human reinforcement learning at the cost of irrational preferences. *Nature Communications*, 9(1), 2018.

[121] Matteo Carandini and David J Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1) :51–62, 2012.

[122] Bolton K.H. Chau, Nils Kolling, Laurence T. Hunt, Mark E. Walton, and Matthew F.S. Rushworth. A neural mechanism underlying failure of optimal choice with multiple alternatives. *Nature Neuroscience*, 17(3) :463–470, 2014.

[123] Chen Hu, Philippe Domenech, and Mathias Pessiglione. Order matters : How covert value updating during sequential option sampling shapes economic preference. *PLoS Computational Biology*, 16(8 August), 2020.

[124] Ian Krajbich, Carrie Armel, and Antonio Rangel. Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, 13(10) :1292–1298, 2010.

[125] Alizée Lopez-Persem, Philippe Domenech, and Mathias Pessiglione. How prior preferences determine decision-making frames and biases in the human brain. *eLife*, 5(NOVEMBER2016) :1–20, 2016.

[126] Samuel J. Gershman and Yael Niv. Learning latent structure : Carving nature at its joints. *Current Opinion in Neurobiology*, 20(2) :251–256, 2010.

[127] Yuan Chang Leong, Angela Radulescu, Reka Daniel, Vivian DeWoskin, and Yael Niv. Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments. *Neuron*, 93(2) :451–463, 2017.

[128] Angela Radulescu, Yeon Soon Shin, and Yael Niv. Human Representation Learning. *Annual Review of Neuroscience*, 44 :253–273, 2021.

[129] Yael Niv, Reka Daniel, Andra Geana, Samuel J. Gershman, Yuan Chang Leong, Angela Radulescu, and Robert C. Wilson. Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, 35(21) :8145–8157, 2015.

[130] Timo Flesch, Keno Juechems, Tsvetomira Dumbalska, Andrew Saxe, and Christopher Summerfield. Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron*, 110(7) :1258–1270.e11, 2022.

[131] Jane X. Wang, Zeb Kurth-Nelson, Dharshan Kumaran, Dhruva Tirumala, Hubert Soyer, Joel Z. Leibo, Demis Hassabis, and Matthew Botvinick. Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, 21(6) :860–868, 2018.

[132] Jill X. O'Reilly. Making predictions in a changing world-inference, uncertainty, and learning. *Frontiers in Neuroscience*, 7(7 JUN) :1–10, 2013.

[133] David C Knill and Alexandre Pouget. The Bayesian brain : the role of uncertainty in neural coding and computation. 27(12), 2004.

[134] Kenji Doya, Shin Ishii, Alexandre Pouget, and Rajesh P N Rao. *Bayesian brain : Probabilistic approaches to neural coding*. MIT press, 2007.

[135] József Fiser and Richard N. Aslin. Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences of the United States of America*, 99(24) :15822–15826, 2002.

[136] József Fiser and Richard N. Aslin. Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, 12(6) :499–504, 2001.

[137] Konrad P. Kording and Daniel M. Wolpert. Bayesian integration in sensorimotor learning. *Nature*, 427(6971) :244–247, 2004.

[138] W Gehring, B Goss, M Coles, D MEyer, and E Donchin. A neural system for error detection. *Psychological Science*, 4(6) :1–6, 1993.

[139] Matthew M. Botvinick, Todd S. Braver, Deanna M. Barch, Cameron S. Carter, and Jonathan D. Cohen. Botvinick et al. 2001 - Conflict monitoring and cognitive control., 2001.

[140] K. A. Hadland, M. F.S. Rushworth, D. Gaffan, and R. E. Passingham. The anterior cingulate and reward-guided selection of actions. *Journal of Neurophysiology*, 89(2) :1161–1164, 2003.

[141] Kenji Matsumoto, Wataru Suzuki, and Keiji Tanaka. Neuronal Correlates of Goal-Based Motor Selection in the Prefrontal Cortex. 301(July) :229–232, 2003.

[142] Céline Amiez, Jean Paul Joseph, and Emmanuel Procyk. Anterior cingulate error-related activity is modulated by predicted reward. *European Journal of Neuroscience*, 21(12) :3447–3452, 2005.

[143] Céline Amiez, Jean-Paul Joseph, and Emmanuel Procyk. Reward encoding in the monkey anterior cingulate cortex. *Cerebral cortex*, 16(7) :1040–1055, 2006.

[144] Joshua W Brown and Todd S Braver. Learned predictions of error likelihood in the anterior cingulate cortex. *Science*, 307(5712) :1118–1121, 2005.

[145] Charles Findling, Nicolas Chopin, and Etienne Koechlin. Imprecise neural computations as a source of adaptive behaviour in volatile environments. *Nature Human Behaviour*, 5(1) :99–112, 2021.

[146] Massimo Silvetti, Ruth Seurinck, and Tom Verguts. Value and prediction error estimation account for volatility effects in ACC : A model-based fMRI study. *Cortex*, 49(6) :1627–1635, 2013.

[147] Rudolf Emil Kalman. Contributions to the theory of optimal control. *Bol. soc. mat. mexicana*, 5(2) :102–119, 1960.

[148] Greg Welch and Gary Bishop. An introduction to the Kalman filter. 1995.

[149] Johan Kwisthout. Most probable explanations in Bayesian networks : Complexity and tractability. *International Journal of Approximate Reasoning*, 52(9) :1452–1469, 2011.

[150] P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel. The Helmholtz machine. *Neural computation*, 7(5) :889–904, 1995.

[151] Christoph Mathys, Jean Daunizeau, Karl J. Friston, and Klaas E. Stephan. A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, 5(MAY) :9, 2011.

[152] Payam Piray and Nathaniel D. Daw. A simple model for learning in volatile environments. *PLoS Computational Biology*, 16(7) :1–26, 2020.

[153] Karl Friston. The free-energy principle : A unified brain theory ? *Nature Reviews Neuroscience*, 11(2) :127–138, 2010.

[154] Karl J. Friston, Jean Daunizeau, and Stefan J. Kiebel. Reinforcement learning or active inference ? *PLoS ONE*, 4(7), 2009.

[155] Johan Kwisthout and Iris van Rooij. Computational Resource Demands of a Predictive Bayesian Brain. *Computational Brain and Behavior*, 3(2) :174–188, 2020.

[156] Johan Kwisthout, Harold Bekkering, and Iris Van Rooij. To be precise, the details don't matter : On predictive processing, precision, and level of detail of predictions. *Brain and cognition*, 112 :84–91, 2017.

[157] Samuel J. Gershman. What does the free energy principle tell us about the brain ? pages 1–10, 2019.

[158] Mel Andrews. The math is not the territory : navigating the free energy principle. *Biology and Philosophy*, 36(3) :1–36, 2021.

[159] Aaron C. Courville, Nathaniel D. Daw, and David S. Touretzky. Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, 10(7) :294–300, 2006.

[160] Samuel J. Gershman. A Unifying Probabilistic View of Associative Learning. *PLoS Computational Biology*, 11(11) :1–20, 2015.

[161] David Marr. *Vision : A computational investigation into the human representation and processing of visual information.* MIT press, 2010.

[162] Sergey Levine. Reinforcement Learning and Control as Probabilistic Inference : Tutorial and Review. 2018.

[163] George Velentzas, Costas Tzafestas, and Mehdi Khamassi. Bridging Computational Neuroscience and Machine Learning on Non-Stationary Multi-Armed Bandits. *bioRxiv*, 2017.

[164] Richard Dearden, Nir Friedman, and Stuart Russell. Bayesian Q-learning. *Proceedings of the National Conference on Artificial Intelligence*, pages 761–768, 1998.

[165] Matthieu Geist and Olivier Pietquin. Kalman temporal differences. *Journal of Artificial Intelligence Research*, 39 :483–532, 2010.

[166] Samuel J. Gershman and Naoshige Uchida. Believing in dopamine. *Nature Reviews Neuroscience*, 20(11) :703–714, 2019.

[167] Clara Kwon Starkweather, Benedicte M Babayan, Naoshige Uchida, and Samuel J Gershman. Dopamine reward prediction errors reflect hidden-state inference across time. *Nature Neuroscience*, 20(4) :581–589, 2017.

[168] Benedicte M. Babayan, Naoshige Uchida, and Samuel J. Gershman. Belief state representation in the dopamine system. *Nature Communications*, 9(1), 2018.

[169] R. J. Herrnstein. Formal Properties of the Matching Law. *Journal of the Experimental Analysis of Behavior*, 21(1) :159–164, 1974.

[170] Gene M Heyman. Is time allocation unconditioned behavior ? In M. Commons, R. Herrnstein, and H. Rachlin, editors, *Quantitative Analyses of Behavior, Vol. 2 : Matching and Maximizing Accounts (Vol.*, pages 459–490. Mass : Ballinger Press, Cambridge, 1982.

[171] John Gibbon, Russell M Church, Stephen Fairhurst, and Alejandro Kacelnik. Scalar Expectancy Theory and Choice Between Delayed Rewards. *Psychological Review*, 95(1) :102–114, 1988.

[172] John Gibbon. Dynamics of time matching : Arousal makes better seem worse. *Psychonomic Bulletin & Review*, 2(2) :208–215, 1995.

[173] C. R. Gallistel, Terence A. Mark, Adam Philip King, and P. E. Latham. The rat approximates an ideal detector of changes in rates of reward : Implications for the law of effect. *Journal of Experimental Psychology : Animal Behavior Processes*, 27(4) :354–372, 2001.

[174] C R Gallistel. Frequency, contingency and the information processing theory of conditioning. In *Frequency Processing and Cognition*, pages 153–172. 2002.

[175] Charles R. Gallistel, Stephen Fairhurst, and Peter Balsam. The learning curve : Implications of a quantitative analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 101(36) :13124–13131, 2004.

[176] Nathaniel D. Daw and Aaron C. Courville. The pigeon as particle filter. *Advances in neural information processing systems*, (20) :369–376, 2007.

[177] Dimitrije Markovic and Stefan J. Kiebel. Comparative analysis of behavioral models for adaptive learning in changing environments. *Frontiers in Computational Neuroscience*, 10(APR), 2016.

[178] Micha Heilbron and Florent Meyniel. Confidence resets reveal hierarchical adaptive learning in humans. *PLoS Computational Biology*, 15(4) :1–24, 2019.

[179] Christopher M Glaze, Joseph W Kable, and Joshua I Gold. Normative evidence accumulation in unpredictable environments. *eLife*, 4 :e08825, aug 2015.

[180] Robert C Wilson, Matthew R Nassar, and Joshua I Gold. Bayesian online learning of the hazard rate in change-point problems, 2010.

[181] Christopher M. Glaze, Alexandre L.S. Filipowicz, Joseph W. Kable, Vijay Balasubramanian, and Joshua I. Gold. A bias-variance trade-off governs individual differences in on-line learning in an unpredictable environment. *Nature Human Behaviour*, 2(3) :213–224, 2018.

[182] Robert C. Wilson, Matthew R. Nassar, and Joshua I. Gold. A Mixture of Delta-Rules Approximation to Bayesian Inference in Change-Point Problems. *PLoS Computational Biology*, 9(7), 2013.

[183] Vincent Moens and Alexandre Zénon. Learning and forgetting using reinforced Bayesian change detection. *PLoS Computational Biology*, 15(4) :1–41, 2019.

[184] Matthew R. Nassar, Robert C. Wilson, Benjamin Heasly, and Joshua I. Gold. An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience*, 30(37) :12366–12378, 2010.

[185] William H. Alexander and Joshua W. Brown. Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, 14(10) :1338–1344, 2011.

[186] Matthew R. Nassar, Joseph T. McGuire, Harrison Ritz, and Joseph W. Kable. Dissociable forms of uncertainty-driven representational change across the human brain. *Journal of Neuroscience*, 39(9) :1688–1698, 2019.

[187] Chang Hao Kao, Ankit N Khambhati, Danielle S Bassett, Matthew R Nassar, Joseph T. McGuire, Joshua I Gold, and Joseph W Kable. Functional brain network reconfiguration during learning in a dynamic environment. *Nature Communications*, 11(1) :1–13, 2020.

[188] Mattias P Karlsson, Dougal G R Tervo, and Alla Y Karpova. Network resets in medial prefrontal cortex mark the onset of behavioral uncertainty. *Science*, 338(6103) :135–139, 2012.

[189] Marwen Belkaid, Elise Bousseyrol, Romain Durand-de Cuttoli, Malou Dongelmans, Etienne K. Duranté, Tarek Ahmed Yahia, Steve Didienne, Bernadette Hanesse, Maxime Come, Alex Mourot, Jérémie Naudé, Olivier Sigaud, and Philippe Faure. Mice adaptively generate choice variability in a deterministic task. *Communications Biology*, 3(1), 2020.

[190] René Quilodran, Marie Rothé, and Emmanuel Procyk. Behavioral Shifts and Action Valuation in the Anterior Cingulate Cortex. *Neuron*, 57(2) :314–325, 2008.

[191] Yoshiya Matsuzaka, Tetsuya Akiyama, Jun Tanji, and Hajime Mushiake. Neuronal activity in the primate dorsomedial prefrontal cortex contributes to strategic selection of response tactics. *Proceedings of the National Academy of Sciences of the United States of America*, 109(12) :4633–4638, 2012.

[192] Philippe Domenech, Sylvain Rheims, and Etienne Koechlin. Neural mechanisms resolving exploitation-exploration dilemmas in the medial prefrontal cortex. *Science*, 369(6507) :eabb0184, aug 2020.

[193] Benjamin Y. Hayden, John M. Pearson, and Michael L. Platt. Neuronal basis of sequential foraging decisions in a patchy environment. *Nature Neuroscience*, 14(7) :933–939, 2011.

[194] Nils Kolling and Jill X. O'Reilly. State-change decisions and dorsomedial prefrontal cortex : the importance of time, 2018.

[195] Elsa F. Fouragnan, Bolton K.H. Chau, Davide Folloni, Nils Kolling, Lennart Verhagen, Miriam Klein-Flügge, Lev Tankelevitch, Georgios K. Papageorgiou, Jean Francois Aubry, Jerome Sallet, and Matthew F.S. Rushworth. The macaque anterior cingulate cortex translates counterfactual choice value into actual behavioral change. *Nature Neuroscience*, 22(5) :797–808, 2019.

[196] Morteza Sarafyazd and Mehrdad Jazayeri. Hierarchical reasoning by neural circuits in the frontal cortex. *Science*, 364(6441) :eaav8911, 2019.

[197] Mattia Rigotti, Omri Barak, Melissa R. Warden, Xiao Jing Wang, Nathaniel D. Daw, Earl K. Miller, and Stefano Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451) :585–590, 2013.

[198] Vinod Venkatraman and Scott A. Huettel. Strategic control in decision-making under uncertainty. *European Journal of Neuroscience*, 35(7) :1075–1082, 2012.

[199] Jérôme Sallet, Rogier B. Mars, Maryann P. Noonan, Franz Xaver Neubert, Saad Jbabdi, Jill X. O'Reilly, Nicola Filippini, Adam G. Thomas, and Matthew F. Rushworth. The organization of dorsal frontal cortex in humans and macaques. *Journal of Neuroscience*, 33(30) :12255–12274, 2013.

[200] Simon B. Eickhoff, Angela R. Laird, Peter T. Fox, Danilo Bzdok, and Lukas Hensel. Functional Segregation of the Human Dorsomedial Prefrontal Cortex, 2016.

[201] Kenji Doya, Kazuyuki Samejima, Ken-ichi Katagiri, and Mitsuo Kawato. Multiple Model-Based Reinforcement Learning. *Neural computation*, 14 :1347–1369, 2002.

[202] Gianluca 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.

[203] Mehdi Khamassi, Louis Emmanuel Martinet, and Agnès Guillot. Combining self-organizing maps with mixtures of experts : Application to an actor-critic model of reinforcement learning in the basal ganglia. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 4095 LNAI :394–405, 2006.

[204] John P. O'Doherty, Sang Wan Lee, and Daniel McNamee. The structure of reinforcement-learning mechanisms in the human brain. *Current Opinion in Behavioral Sciences*, 1 :94–100, 2015.

[205] John P. O'Doherty, Sangwan Lee, Reza Tadayonnejad, Jeff Cockburn, Kyo Iigaya, and Caroline J. Charpentier. Why and how the brain weights contributions from a mixture of experts. *Neuroscience and Biobehavioral Reviews*, 123 :14–23, 2021.

[206] 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.

[207] 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.

[208] Jan Gläscher, Nathaniel Daw, Peter Dayan, and John P O'Doherty. States versus rewards : dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66(4) :585–595, 2010.

[209] 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(OCTOBER 2012) :1–19, 2012.

[210] Kenji Doya, Hiroyuki Nakahara, Raju S Bapi, and Okihide Hikosaka. Multiple Representations and Algorithms for Sequence Learning. pages 3–5, 1997.

[211] 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.

[212] Bruno Miranda, W. M. Nishantha Malalasekera, Timothy E. Behrens, Peter Dayan, and Steven W. Kennerley. Combined model-free and model-sensitive reinforcement learning in non-human primates. *PLoS Computational Biology*, 16(6) :1–25, 2020.

[213] Erwan Renaudo, Benoît Girard, Raja Chatila, and Mehdi Khamassi. Which criteria for autonomously shifting between goal-directed and habitual behaviors in robots ? In *5th Joint International Conference on Development and Learning and Epigenetic Robotics, ICDL-EpiRob 2015*, pages 254–260, 2015.

[214] Rémi Dromnell, Erwan Renaudo, Guillaume Pourcel, Raja Chatila Benoît, and Girard Mehdi 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. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12413 LNAI :68–79, 2021.

[215] Lee SW, Shimojo S, and O'Doherty JP. Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, 81(3) :687–699, 2014.

[216] Mehdi Keramati, Amir Dezfouli, and Payam Piray. Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Computational Biology*, 7(5), 2011.

[217] Laurent Dollé, Denis Sheynikhovich, Benoît Girard, Balázs Ujfalussy, Ricardo Chavarriaga, and Agnès Guillot. Analyzing interactions between cue-guided and place-based navigation with a computational model of action selection : Influence of sensory cues and training. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6226 LNAI, pages 335–346, 2010.

[218] Marios C Panayi, Mehdi Khamassi, and Simon Killcross. The rodent lateral orbitofrontal cortex as an arbitrator selecting between model-based and model-free learning systems. *Behavioral Neuroscience*, 135(2) :226–244, 2021.

[219] 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) :1–35, 2018.

[220] James C Houk and James L Adams. 13 A model of how the basal ganglia generate and use neural signals that. *Models of information processing in the basal ganglia*, page 249, 1995.

[221] Brenda Milner. Effects of different brain lesions on card sorting. *Archives of neurology*, 9(1) :90, 1963.

[222] Earl K Miller and Jonathan D Cohen. An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, 24(1) :167–202, 2001.

[223] MICHAEL N. SHADLEN and AND WILLIAM T. NEWSOME. Neural Basis of a Perceptual Decision in the Parietal Cortex (Area LIP) of the Rhesus Monkey. *Journal of Neurophysiology*, 86 :1916–1936, 2001.

[224] Aldo Genovesio, Peter J Brasted, and Steven P Wise. Representation of future and previous spatial goals by separate neural populations in prefrontal cortex.

The Journal of neuroscience : the official journal of the Society for Neuroscience, 26(27) :7305–7316, 2006.

[225] Carlo Reverberi, Kai Görgen, and John Dylan Haynes. Compositionality of rule representations in human prefrontal cortex. *Cerebral Cortex*, 22(6) :1237–1246, 2012.

[226] Gordon D Logan and Robert D Gordon. Executive control of visual attention in dual-task situations. *Psychological review*, 108(2) :393, 2001.

[227] Nico U.F. Dosenbach, Kristina M. Visscher, Erica D. Palmer, Francis M. Miezin, Kristin K. Wenger, Hyunseon C. Kang, E. Darcy Burgund, Ansley L. Grimes, Bradley L. Schlaggar, and Steven E. Petersen. A Core System for the Implementation of Task Sets. *Neuron*, 50(5) :799–812, 2006.

[228] Katsuyuki Sakai. Task set and prefrontal cortex. *Annu. Rev. Neurosci.*, 31 :219–245, 2008.

[229] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4) :594–611, 2006.

[230] Sang Wan Lee, John P. O'Doherty, and Shinsuke Shimojo. Neural Computations Mediating One-Shot Learning in the Human Brain. *PLoS Biology*, 13(4) :1–36, 2015.

[231] Brenden M Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua B Tenenbaum. One shot learning of simple visual concepts. *In Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, 2011.

[232] Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266) :1332–1338, 2015.

[233] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Sharing clusters among related groups : Hierarchical dirichlet processes. In *Advances in Neural Information Processing Systems*, number 1, 2005.

[234] Thomas L. Griffiths, Kevin R. Canini, Adam N. Sanborn, and Daniel J. Navarro. Unifying rational models of categorization via the hierarchical Dirichlet process. *Proceedings of the 29th annual conference of the cognitive science society*, page 323328, 2007.

[235] Adam N Sanborn, Thomas L Griffiths, and Daniel J Navarro. Rational approximations to rational models : alternative algorithms for category learning. *Psychological review*, 117(4) :1144, 2010.

[236] Johan Kwisthout, Todd Wareham, and Iris Van Rooij. Bayesian intractability is not an ailment that approximation can cure. *Cogn. Sci.*, 35(5) :779–784, 2011.

[237] Charles Kemp, Noah D Goodman, and Joshua B Tenenbaum. Learning causal schemata. *Proceedings of the 29th annual conference of the cognitive science society*, pages 389–394, 2007.

[238] Amy Perfors, Joshua B Tenenbaum, and Terry Regier. The learnability of abstract syntactic principles. *Cognition*, 118(3) :306–338, 2011.

[239] Daniel Durstewitz, Nicole M. Vittoz, Stan B. Floresco, and Jeremy K. Seamans. Abrupt transitions between prefrontal neural ensemble states accompany behavioral transitions during rule learning. *Neuron*, 66(3) :438–448, 2010.

[240] A. Saez, M. Rigotti, S. Ostojic, S. Fusi, and C. D. Salzman. Abstract Context Representations in Primate Amygdala and Prefrontal Cortex. *Neuron*, 87(4) :869–881, 2015.

[241] Maïlys C.M. Faraut, Emmanuel Procyk, and Charles R.E. Wilson. Learning to learn about uncertain feedback. *Learning and Memory*, 23(2) :90–98, 2016.

[242] Nicolas W. Schuck, Robert Gaschler, Dorit Wenke, Jakob Heinzle, Peter A. Frensch, John Dylan Haynes, and Carlo Reverberi. Medial prefrontal cortex predicts internally driven strategy shifts. *Neuron*, 86(1) :331–340, 2015.

[243] David Badre, Andrew S. Kayser, and Mark D'Esposito. Frontal cortex and the discovery of abstract action rules. *Neuron*, 66(2) :315–326, 2010.

[244] Etienne Koechlin and Christopher Summerfield. An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences*, 11(6) :229–235, 2007.

[245] David Badre and Mark D'Esposito. Is the rostro-caudal axis of the frontal lobe hierarchical ? *Nature Reviews Neuroscience*, 10(9) :659–669, 2009.

[246] Etiene Koechlin and Nicolas Franck. The Architecture of Cognitive Control in the Human Prefrontal Cortex Related papers Organizat ion of Cognit ive Cont rol Within the Lateral Prefrontal Cortex in Schizophrenia. *Science*, 302 :1181–1185, 2003.

[247] Anne G E Collins and Michael J Frank. How much of reinforcement learning is working memory , not reinforcement learning ? A behavioral , computational , and neurogenetic analysis. *European Journal of Neuroscience*, 35(December 2011) :1024–1035, 2012.

[248] S. Ritter, J. X. Wang, Z. Kurth-Nelson, and M. Botvinick. Episodic Control as Meta-Reinforcement Learning. *bioRxiv*, pages 948–953, 2018.

[249] Anne Collins and Etienne Koechlin. Reasoning, learning, and creativity : Frontal lobe function and human decision-making. *PLoS Biology*, 10(3), 2012.

[250] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *34th International Conference on Machine Learning, ICML 2017*, 3 :1856–1868, 2017.

[251] Jan Humplik, Alexandre Galashov, Leonard Hasenclever, Pedro A. Ortega, Yee Whye Teh, and Nicolas Heess. Meta reinforcement learning as task inference. 2019.

[252] Jane X. Wang. Meta-learning in natural and artificial intelligence. *Current Opinion in Behavioral Sciences*, 38 :90–95, 2021.

[253] Camillo Padoa-Schioppa and John A. Assad. Neurons in the orbitofrontal cortex encode economic value. *Nature*, 441(7090) :223–226, 2006.

[254] Benedetto De Martino, Stephen M Fleming, Neil Garrett, and Raymond J Dolan. Confidence in value-based choice. *Nature Neuroscience*, 16(1) :105–110, 2013.

[255] Maël Lebreton, Raphaëlle Abitbol, Jean Daunizeau, and Mathias Pessiglione. Automatic integration of confidence in the brain valuation signal. *Nature neuroscience*, 18(8) :1159–1167, 2015.

[256] Nils Kolling, Timothy E.J. Behrens, Rogier B. Mars, and Matthew F.S. Rushworth. Neural mechanisms of foraging. *Science*, 335(6077) :95–98, 2012.

[257] Chris J Mitchell, Jan De Houwer, and Peter F Lovibond. The propositional nature of human associative learning. *Behavioral and Brain Sciences*, 32(2) :183–198, 2009.

[258] Alireza Soltani and Alicia Izquierdo. Adaptive learning under expected and unexpected uncertainty. *Nature Reviews Neuroscience*, 20(10) :635–644, 2019.

[259] Aurelio Cortese, Asuka Yamamoto, Maryam Hashemzadeh, Pradyumna Sepulveda, Mitsuo Kawato, and Benedetto De Martino. Value signals guide abstraction during learning. *eLife*, 10 :1–27, 2021.

[260] Shiva Farashahi, Jane Xu, Shih Wei Wu, and Alireza Soltani. Learning arbitrary stimulus-reward associations for naturalistic stimuli involves transition from learning about features to learning about objects. *Cognition*, 205 :1–34, 2020.

[261] Alan N. Hampton, Peter Bossaerts, and John P. O'Doherty. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *Journal of Neuroscience*, 26(32) :8360–8367, 2006.

[262] Mathias Pessiglione and Jean Daunizeau. Bridging across functional models : The OFC as a value-making neural network. *Behavioral Neuroscience*, 135(2) :277–290, 2021.

[263] Amos Tversky and Daniel Kahneman. Rational choice and the framing of decisions. In *Multiple criteria decision making and risk analysis using microcomputers*, pages 81–126. Springer, 1989.

[264] Benedetto De Martino, Dharshan Kumaran, Ben Seymour, and Raymond J. Dolan. Frames, biases and rational decision-making in the human brain. *Science*, 313(5787) :684–687, 2006.

[265] Maël Lebreton, Karin Bacily, Stefano Palminteri, and Jan B. Engelmann. Contextual influence on confidence judgments in human reinforcement learning. *PLoS Computational Biology*, 15(4) :1–27, 2019.

[266] Brian Silston, Toby Wise, Song Qi, Xin Sui, Peter Dayan, and Dean Mobbs. Neural encoding of perceived patch value during competitive and hazardous virtual foraging. *Nature Communications*, 12(1) :1–11, 2021.

[267] Bruno B Averbeck. Theory of Choice in Bandit, Information Sampling and Foraging Tasks. *PLoS Computational Biology*, 11(3) :1–28, 2015.

[268] Aurélien Weiss, Valérian Chambon, Junseok K. Lee, Jan Drugowitsch, and Valentin Wyart. Interacting with volatile environments stabilizes hidden-state inference and its brain signatures. *Nature Communications*, 12(1) :1–56, 2021.

[269] Marion Rouault, Aurélien Weiss, Junseok K Lee, Jan Drugowitsch, Valerian Chambon, and Valentin Wyart. Controllability boosts neural and cognitive signatures of changes-of-mind in uncertain environments. *eLife*, 11 :1–28, 2022.

[270] Bradley B Doll, W Jake Jacobs, Alan G Sanfey, and Michael J Frank. Instructional control of reinforcement learning : A behavioral and neurocomputational investigation. *Brain Research*, 1299 :74–94, 2009.

[271] Anne Marike Schiffer, Kayla Siletti, Florian Waszak, and Nick Yeung. Adaptive behaviour and feedback processing integrate experience and instruction in reinforcement learning. *NeuroImage*, 146 :626–641, 2017.

[272] Carolina Feher da Silva and Todd A. Hare. Humans primarily use model-based inference in the two-stage task. *Nature Human Behaviour*, 4(10) :1053–1066, 2020.

[273] Pedro Castro-Rodrigues, Thomas Akam, Ivar Snorasson, Marta Camacho, Vitor Paixão, Ana Maia, J Bernardo Barahona-Corrêa, Peter Dayan, H Blair Simpson, and Rui M Costa. Explicit knowledge of task structure is a primary determinant of human model-based action. *Nature Human Behaviour*, pages 1–16, 2022.

[274] Wouter Kool, Fiery A. Cushman, and Samuel J. Gershman. When Does Model-Based Control Pay Off ? *PLoS Computational Biology*, 12(8) :1–34, 2016.

[275] Keno Juechems and Christopher Summerfield. Where Does Value Come From ? *Trends in Cognitive Sciences*, 23(10) :836–850, 2019.

[276] Mehdi Keramati and Boris Gutkin. Homeostatic reinforcement learning for integrating reward collection and physiological stability. *eLife*, 3 :1–26, 2014.

[277] Wolfram Schultz, Léon Tremblay, and Jeffrey R. Hollerman. Reward processing in primate orbitofrontal cortex and basal ganglia. *Cerebral Cortex*, 10(3) :272–283, 2000.

[278] Benjamin Y Hayden and Yael Niv. The case against economic values in the orbitofrontal cortex (or anywhere else in the brain). *Behavioral Neuroscience*, 135(2) :192–201, 2021.

[279] Justin M. Fine and Benjamin Y. Hayden. The whole prefrontal cortex is premotor cortex. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 377(1844) :20200524, 2022.

[280] David J-N Maisson, Tyler V Cash-Padgett, Maya Z Wang, Benjamin Y Hayden, Sarah R Heilbronner, and Jan Zimmermann. Choice-relevant information transformation along a ventrodorsal axis in the medial prefrontal cortex. *Nature communications*, 12(1) :1–14, 2021.

[281] Raunak Basu, Robert Gebauer, Tim Herfurth, Simon Kolb, Zahra Golipour, Tatjana Tchumatchenko, and Hiroshi T Ito. The orbitofrontal cortex maps future navigational goals. *Nature*, 599(7885) :449–452, 2021.

[282] Sridhar Mahadevan and Mauro Maggioni. Proto-value functions : A laplacian framework for learning representation and control in markov decision processes. *Journal of Machine Learning Research*, 8(74) :2169–2231, 2007.

[283] Kimberly L. Stachenfeld, Matthew M. Botvinick, and Samuel J. Gershman. The hippocampus as a predictive map. *Nature Neuroscience*, 20(11) :1643–1653, 2017.

[284] James C.R. Whittington, Timothy H. Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy E.J. Behrens. The Tolman-Eichenbaum Machine : Unifying Space and Relational Memory through Generalization in the Hippocampal Formation. *Cell*, 183(5) :1249–1263.e23, 2020.

[285] Edward C Tolman. Cognitive maps in rats and men. *Psychological review*, 55(4) :189, 1948.

[286] John O'keefe and Lynn Nadel. Précis of O'Keefe & Nadel's The hippocampus as a cognitive map. *Behavioral and Brain Sciences*, 2(4) :487–494, 1979.

[287] A. David Redish and David S. Touretzky. Cognitive maps beyond the hippocampus. *Hippocampus*, 7(1) :15–35, 1997.

[288] Howard Eichenbaum and Neal J Cohen. Can We Reconcile the Declarative Memory and Spatial Navigation Views on Hippocampal Function ?, 2014.

[289] Dharshan Kumaran, Jennifer J Summerfield, Demis Hassabis, and Eleanor A Maguire. Tracking the Emergence of Conceptual Knowledge during Human Decision Making. *Neuron*, 63(6) :889–901, 2009.

[290] Romain Cazé, Mehdi Khamassi, Lise Aubin, and Benoît Girard. Hippocampal replays under the scrutiny of reinforcement learning models, 2018.

[291] Yunzhe Liu, Marcelo G. Mattar, Timothy E.J. Behrens, Nathaniel D. Daw, and Raymond J. Dolan. Experience replay is associated with efficient nonlocal learning. *Science*, 372(6544), 2021.

[292] Anna C Schapiro, Nicholas B. Turk-Browne, Kenneth A Norman, and Matthew M Botvinick. Statistical learning of temporal community structure in the hippocampus. *Hippocampus*, 26(1) :3–8, 2016.

[293] Silvia Bernardi, Marcus K Benna, Mattia Rigotti, Jérôme Munuera, Stefano Fusi, and C Daniel Salzman. The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell*, 183(4) :954–967.e21, 2020.

[294] Seongmin A Park, Douglas S Miller, Hamed Nili, Charan Ranganath, and Erie D Boorman. Map Making : Constructing, Combining, and Inferring on Abstract Cognitive Maps. *Neuron*, 107(6) :1226–1238.e8, 2020.

[295] Aldo Genovesio, Steven P Wise, and Richard E Passingham. Prefrontal–parietal function : from foraging to foresight. *Trends in cognitive sciences*, 18(2) :72–81, 2014.

[296] Christopher Summerfield, Fabrice Luyckx, and Hannah Sheahan. Structure learning and the posterior parietal cortex. *Progress in Neurobiology*, 184 :101717, 2020.

[297] Dedre Gentner. Structure-Mapping : A Theoretical Framework for Analogy. *Readings in Cognitive Science : A Perspective from Psychology and Artificial Intelligence*, pages 303–310, 1983.

[298] Ronald B. Dekker, Fabian Otto, and Christopher Summerfield. Curriculum learning for human compositional generalization. *Proceedings of the National Academy of Sciences of the United States of America*, 119(41) :1–12, 2022.

[299] Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40(2012) :1–58, 2017.

[300] Fabrice Luyckx, Hamed Nili, Bernhard Spitzer, and Christopher Summerfield. Neural structure mapping in human probabilistic reward learning. *eLife*, 8 :1–19, 2019.

[301] Hannah Sheahan, Fabrice Luyckx, Stephanie Nelli, Clemens Teupe, and Christopher Summerfield. Neural state space alignment for magnitude generalization in humans and recurrent networks. *Neuron*, 109(7) :1214–1226.e8, 2021.

[302] Pedro A. Tsividis, Joao Loula, Jake Burga, Nathan Foss, Andres Campero, Thomas Pouncy, Samuel J. Gershman, and Joshua B. Tenenbaum. Human-Level Reinforcement Learning through Theory-Based Modeling, Exploration, and Planning. *arXiv preprint arXiv*, pages 1–67, 2021.

[303] Tom Schaul. A video game description language for model-based or interactive learning. *IEEE Conference on Computatonal Intelligence and Games, CIG*, 2013.

[304] Momchil S. Tomov, Pedro A. Tsividis, Thomas Pouncy, Joshua B. Tenenbaum, and Samuel J. Gershman. The Neural Architecture of Theory-based Reinforcement Learning. *bioRxiv*, page 2022.06.14.496001, 2022.

[305] Iris Van Rooij, Mark Blokpoel, Johan Kwisthout, and Todd Wareham. *Cognition and intractability : A guide to classical and parameterized complexity analysis*. Cambridge University Press, 2019.

[306] Mark Blokpoel, Todd Wareham, Pim Haselager, Ivan Toni, and Iris Van Rooij. Deep Analogical Inference as the Origin of Hypotheses. *Journal of Problem Solving*, 11, 2018.

[307] Nicholas A Roy, Ji Hyun Bak, The International Brain Laboratory, Athena Akrami, Carlos D Brody, and Jonathan W Pillow. Extracting the dynamics of behavior in sensory decision-making experiments. *Neuron*, 109(4) :597–610, 2021.

[308] Zoe C Ashwood, Nicholas A Roy, Iris R Stone, Anne E Urai, Anne K Churchland, Alexandre Pouget, and Jonathan W Pillow. Mice alternate between discrete strategies during perceptual decision-making. *Nature Neuroscience*, 25(2) :201–212, 2022.

[309] Matthew M. Botvinick, Yael Niv, and Andrew C. Barto. Hierarchically organized behavior and its neural foundations : A reinforcement learning perspective. *Cognition*, 113(3) :262–280, 2009.

[310] Frederic M. Stoll, Vincent Fontanier, and Emmanuel Procyk. Specific frontal neural dynamics contribute to decisions to check. *Nature Communications*, 7(May) :1–14, 2016.

[311] Sarah R. Heilbronner and Benjamin Y. Hayden. Dorsal Anterior Cingulate Cortex : A Bottom-Up View. *Annual Review of Neuroscience*, 39(April) :149–170, 2016.

[312] Verity J. Brown and Eric M. Bowman. Rodent models of prefrontal cortical function. *Trends in Neurosciences*, 25(7) :340–343, 2002.

[313] Sarah R Heilbronner, Jose Rodriguez-Romaguera, Gregory J Quirk, Henk J Groenewegen, and Suzanne N Haber. Circuit-based corticostriatal homologies between rat and primate. *Biological psychiatry*, 80(7) :509–521, 2016.

[314] Mark Laubach, Linda M. Amarante, Kyra Swanson, and Samantha R. White. What, if anything, is rodent prefrontal cortex ? *eNeuro*, 5(5), 2018.

[315] Bernard W Balleine and John P. O'Doherty. Human and rodent homologies in action control : Corticostriatal determinants of goal-directed and habitual action, 2010.

[316] Daniel G Woolley, Annelies Laeremans, Ilse Gantois, Dante Mantini, Ben Vermaercke, Hans P. Op De Beeck, Stephan P. Swinnen, Nicole Wenderoth, Lutgarde Arckens, and Rudi D'Hooge. Homologous involvement of striatum and prefrontal cortex in rodent and human water maze learning. *Proceedings of the National Academy of Sciences of the United States of America*, 110(8) :3131–3136, 2013.

[317] Marine Euvrard. Stratégies d'adaptation de la souris face à un environnement volatil, 2021.

[318] Louis Lefebvre, Nektaria Nicolakakis, and Denis Boire. Tools and brains in birds. *Behaviour*, 139(7) :939–973, 2002.

[319] Tore Slagsvold and Karen L. Wiebe. Social learning in birds and its role in shaping a foraging niche, 2011.

[320] Lena Veit, Lucas Y. Tian, Christian J.Monroy Hernandez, and Michael S. Brainard. Songbirds can learn flexible contextual control over syllable sequencing. *eLife*, 10 :1–19, 2021.

[321] B. F. Skinner. Are theories of learning necessary ? *Psychological Review*, 57(4) :193–216, 1950.

[322] Alexander A. Aarts, Joanna E. Anderson, Christopher J. Anderson, Peter R. Attridge, Angela Attwood, Jordan Axt, Molly Babel, Štěpán Bahník, Erica Baranski, Michael Barnett-Cowan, Elizabeth Bartmess, Jennifer Beer, Raoul Bell, Heather Bentley, Leah Beyan, Grace Binion, Denny Borsboom, Annick Bosch, Frank A. Bosco, Sara D. Bowman, Mark J. Brandt, Erin Braswell, Hilmar Brohmer, Benjamin T. Brown, Kristina Brown, Jovita Brüning, Ann Calhoun-Sauls, Shannon P. Callahan, Elizabeth Chagnon, Jesse Chandler, Christopher R. Chartier, Felix Cheung, Cody D. Christopherson, Linda Cillessen, Russ Clay, Hayley Cleary, Mark D. Cloud, Michael Conn, Johanna Cohoon, Simon Columbus,

Andreas Cordes, Giulio Costantini, Leslie D.Cramblet Alvarez, Ed Cremata, Jan Crusius, Jamie DeCoster, Michelle A. DeGaetano, Nicolás Delia Penna, Bobby Den Bezemer, Marie K. Deserno, Olivia Devitt, Laura Dewitte, David G. Dobolyi, Geneva T. Dodson, M. Brent Donnellan, Ryan Donohue, Rebecca A. Dore, Angela Dorrough, Anna Dreber, Michelle Dugas, Elizabeth W. Dunn, Kayleigh Easey, Sylvia Eboigbe, Casey Eggleston, Jo Embley, Sacha Epskamp, Timothy M. Errington, Vivien Estel, Frank J. Farach, Jenelle Feather, Anna Fedor, Belén Fernández-Castilla, Susann Fiedler, James G. Field, Stanka A. Fitneva, Taru Flagan, Amanda L. Forest, Eskil Forsell, Joshua D. Foster, Michael C. Frank, Rebecca S. Frazier, Heather Fuchs, Philip Gable, Jeff Galak, Elisa Maria Galliani, Anup Gampa, Sara Garcia, Douglas Gazarian, Elizabeth Gilbert, Roger Giner-Sorolla, Andreas Glöckner, Lars Goellner, Jin X. Goh, Rebecca Goldberg, Patrick T. Goodbourn, Shauna Gordon-McKeon, Bryan Gorges, Jessie Gorges, Justin Goss, Jesse Graham, James A. Grange, Jeremy Gray, Chris Hartgerink, Joshua Hartshorne, Fred Hasselman, Timothy Hayes, Emma Heikensten, Felix Henninger, John Hodsoll, Taylor Holubar, Gea Hoogendoorn, Denise J. Humphries, Cathy O.Y. Hung, Nathali Immelman, Vanessa C. Irsik, Georg Jahn, Frank Jäkel, Marc Jekel, Magnus Johannesson, Larissa G. Johnson, David J. Johnson, Kate M. Johnson, William J. Johnston, Kai Jonas, Jennifer A. Joy-Gaba, Heather Barry Kappes, Kim Kelso, Mallory C. Kidwell, Seung Kyung Kim, Matthew Kirkhart, Bennett Kleinberg, Goran Knežević, Franziska Maria Kolorz, Jolanda J. Kossakowski, Robert Wilhelm Krause, Job Krijnen, Tim Kuhlmann, Yoram K. Kunkels, Megan M. Kyc, Calvin K. Lai, Aamir Laique, Daniël Lakens, Kristin A. Lane, Bethany Lassetter, Ljiljana B. Lazarević, Etienne P. Le Bel, Key Jung Lee, Minha Lee, Kristi Lemm, Carmel A. Levitan, Melissa Lewis, Lin Lin, Stephanie Lin, Matthias Lippold, Darren Loureiro, Ilse Luteijn, Sean MacKinnon, Heather N. Mainard, Denise C. Marigold, Daniel P. Martin, Tylar Martinez, E. J. Masicampo, Josh Matacotta, Maya Mathur, Michael May, Nicole Mechin, Pranjal Mehta, Johannes Meixner, Alissa Melinger, Jeremy K. Miller, Mallorie Miller, Katherine Moore, Marcus Möschl, Matt Motyl, Stephanie M. Müller, Marcus Munafo, Koen I. Neijenhuijs, Taylor Nervi, Gandalf Nicolas, Gustav Nilsonne, Brian A. Nosek, Michèle B. Nuijten, Catherine Olsson, Colleen Osborne, Lutz Ostkamp, Misha Pavel, Ian S. Penton-Voak, Olivia Perna, Cyril Pernet, Marco Perugini, R. Nathan Pipitone, Michael Pitts, Franziska Plessow, Jason M. Prenoveau, Rima Maria Rahal, Kate A. Ratliff, David Reinhard, Frank Renkewitz, Ashley A. Ricker, Anastasia Rigney, Andrew M. Rivers, Mark Roebke, Abraham M. Rutchick, Robert S. Ryan, Onur Sahin, Anondah Saide, Gillian M. Sandstrom, David Santos, Rebecca Saxe, René Schlegelmilch, Kathleen Schmidt, Sabine Scholz, Larissa Seibel, Dylan Faulkner Selterman, Samuel Shaki, William B. Simpson, H. Colleen Sinclair, Jeanine L.M. Skorinko, Agnieszka Slowik, Joel S. Snyder, Courtney Soderberg, Carina Sonnleitner, Nick Spencer, Jeffrey R. Spies, Sara Steegen, Stefan Stieger, Nina Strohminger, Gavin B. Sullivan, Thomas Talhelm, Megan Tapia, Anniek Te Dorsthorst, Manuela Thomae, Sarah L. Thomas, Pia Tio, Frits Traets, Steve Tsang,

Francis Tuerlinckx, Paul Turchan, Milan Valášek, Anna E. Van't Veer, Robbie Van Aert, Marcel Van Assen, Riet Van Bork, Mathijs Van De Ven, Don Van Den Bergh, Marije Van Der Hulst, Roel Van Dooren, Johnny Van Doorn, Daan R. Van Renswoude, Hedderik Van Rijn, Wolf Vanpaemel, Alejandro Vásquez Echeverría, Melissa Vazquez, Natalia Velez, Marieke Vermue, Mark Verschoor, Michelangelo Vianello, Martin Voracek, Gina Vuu, Eric Jan Wagenmakers, Joanneke Weerdmeester, Ashlee Welsh, Erin C. Westgate, Joeri Wissink, Michael Wood, Andy Woods, Emily Wright, Sining Wu, Marcel Zeelenberg, and Kellylynn Zuni. Estimating the reproducibility of psychological science. *Science*, 349(6251), 2015.

[323] Berna Devezer, Luis G Nardin, Bert Baumgaertner, and Erkan Ozge Buzbas. Scientific discovery in a model-centric framework : Reproducibility, innovation, and epistemic diversity. *PloS one*, 14(5) :e0216125, 2019.

[324] Berna Devezer, Danielle J Navarro, Joachim Vandekerckhove, and Erkan Ozge Buzbas. The case for formal methodology in scientific reform. *Royal Society Open Science*, 8(3), 2021.

[325] Patricia Rich, Ronald de Haan, Todd Wareham, and Iris van Rooij. How hard is cognitive science ? In *Proceedings of the annual meeting of the cognitive science society*, volume 43, 2021.